Introducción al uso del Generador Estocástico de series sintéticas

Alessio Bocco (boccoalessio@gmail.com)
Daniel Bonhaure (danielbonhaure@gmail.com)
Guillermo Podestá (gpodesta@rsmas.miami.edu)

15 de December de 2020

Contents

1 Introducción **1**

2 Instalación de paquetes necesarios **2**

3 Creación de directorios **2**

4 Generación de series sintéticas para una sola estación meteorológica **2**

4.1 Crear archivos de entrada 2

4.2 Ajuste de los modelos estadísticos 5

5 Generación de series sintéticas **9**

5.1 Resultado de la generación 12

6 Diagnósticos **12**

6.1 Diagnósticos de precipitación 13

6.2 Diagnósticos de temperatura máxima 26

6.3 Diagnósticos de temperatura mínima 33

6.4 Diagnósticos de auxiliares 42

1 Introducción

El presente documento tiene como propósito proveer ejemplos de aplicación del Generador Estocástico de datos climáticos desarrollado por el SISSA. Este software es una herramienta clave para el análisis probabilista del riesgo de sequías pero por su flexibilidad puede ser usado en una multitud de aplicaciones.

Esta guía contiene solamente ejemplos de *algunas* de las posibilidades y productos derivados del generador. Es un complemento de la presentación y solo se desarrollarán y replicarán algunos de los contenidos mostrados durante la misma.

2 Instalación de paquetes necesarios

Se comprueba que estén instalados los paquetes necesarios. De no ser así, se instalarán automáticamente.

3 Creación de directorios

En este paso nos aseguramos de crear los directorios que contendrán los datos de entrada para el ajuste del modelo y los resultados.

- /input_data: aquí se guardarán los datos meteorológicos y los metadatos de las estaciones
- /output_data: aquí se guardarán los resultados de la simulación

Si estos directorios no existen, se crearán.

4 Generación de series sintéticas para una sola estación meteorológica

En este ejemplo veremos como se generan series sintéticas para una sola estación meteorológica condicionada por variables trimestrales. Esta configuración permite que las series generadas copien los cambios observadas en el registro histórico.

4.1 Crear archivos de entrada

El primer paso consiste en generar los set de datos de entrada que se descargaron al momento de instalar el paquete del generador estocástico. Estos datos son sólo a título demostrativo, si el usuario desea correr el modelo con sus propios datos deberá cambiar los objetos que se generarán en esta sección por los suyos y colocarlos en la carpeta `input_data`.

Los archivos necesarios son:

- stations.csv
- climate.csv

Los datos meteorológicos se dividen en dos archivos separados: `stations.csv` y `climate.csv`. Los nombres de los mismos no deben ser necesariamente iguales a los usados aquí.

Los metadatos de las estaciones se alojan en el archivo `stations.csv`. Este archivo contiene la información de las estaciones meteorológicas que serán usadas en el ajuste del modelo. Las variables que deben ser incluidas en la tabla son:

- station_id: número unívoco para cada estación meteorológica. La variable debe ser de tipo *integer*
- latitude: latitud en grados decimales. La variable debe ser de tipo *double*
- longitude: longitud en grados decimales. La variable debe ser de tipo *double*

La tabla puede tener más variables pero sólo se necesitan las anteriores.

A continuación se muestran la primera fila del dataset y los tipo de datos de cada una de las variables.

```
# Vista de los metadatos de la estación
knitr::kable(stations[1,])
```

x	y	station_id	nombre	lat_dec	lon_dec	elev	pais_id
5001614	6256841	87448	Villa Reynolds Aero	-33.7181	-65.3737	486	AR

El objeto **stations** debe ser convertido de *tibble* a *sf*. El sistema de referencia espacial debe ser planar. No es necesario un sistema de referencia espacial en particular, solamente las coordenadas deben estar expresadas en metros.

```
# Convertimos el objeto stations a sf y se convierte su proyección de WGS 1984 a
# POSGAR Argentina Faja 5.
stations %<>%
  sf::st_as_sf(coords = c("lon_dec", "lat_dec"), crs =  4326) %>%
  sf::st_transform(crs = 22185)
```

La información climática se aloja en el archivo `climate.csv`. Este archivo contiene los datos de las estaciones meteorológicas que serán usadas en el ajuste del modelo. Las variables que deben ser incluidas en la tabla son:

- `date`: fecha del dato. La variable debe ser de tipo *date*
- `station_id`: número unívoco para cada estación meteorológica. La variable debe ser de tipo *integer*
- `prcp`: datos diarios de precipitación La variable debe ser de tipo *double*
- `tmax`: datos diarios de temperatura máxima. La variable debe ser de tipo *double*
- `tmin`: datos diarios de temperatura mínima. La variable debe ser de tipo *double*

A continuación se muestran las primeras cinco filas del dataset y los tipos de datos de cada una de las variables.

```
knitr::kable(climate[1:10,])
```

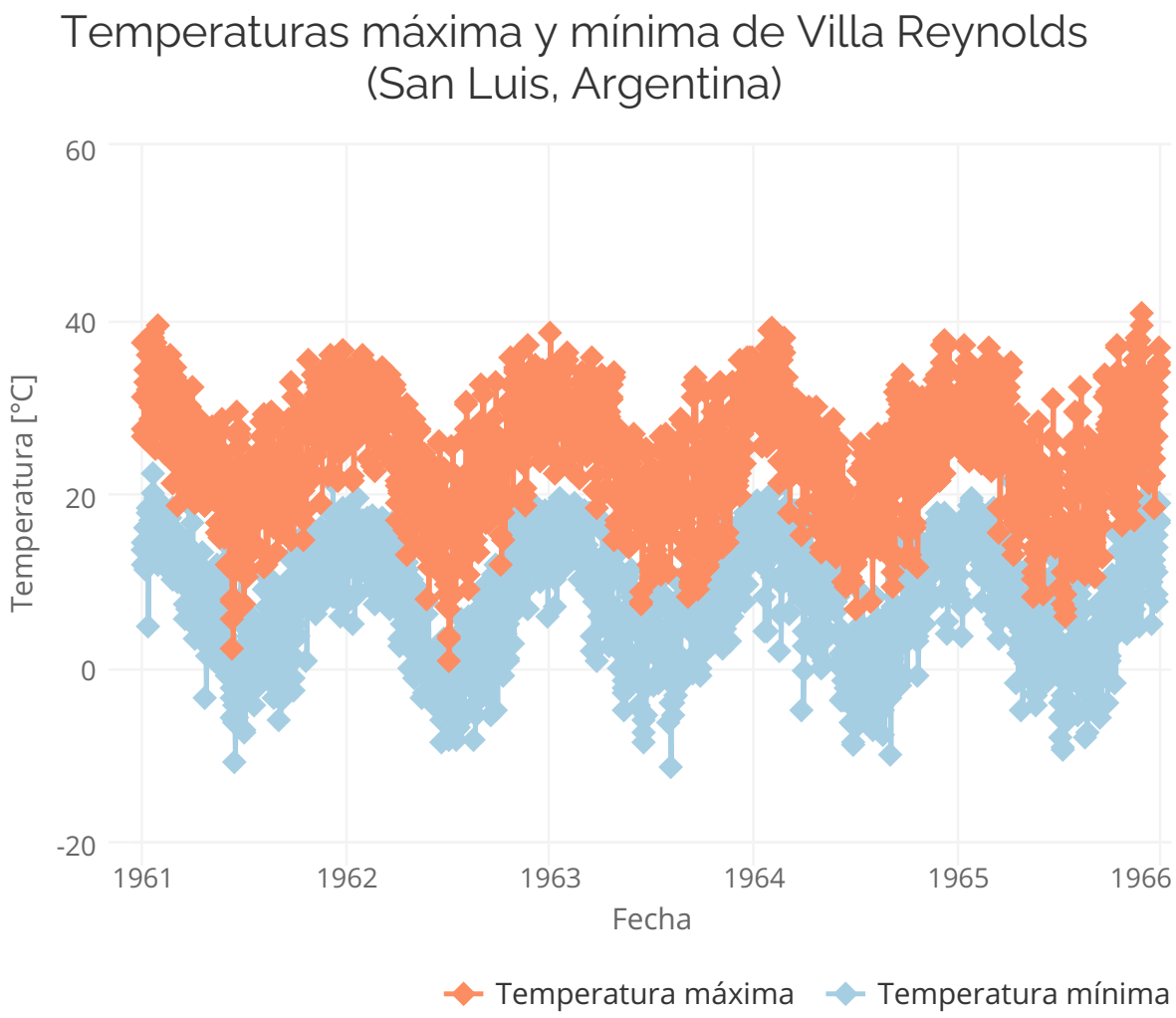
date	station_id	tmax	tmin	prcp
1961-01-01	87448	37.4	13.5	0.6
1961-01-02	87448	27.4	14.3	23.9
1961-01-03	87448	26.6	13.5	0.0
1961-01-04	87448	31.0	11.7	6.0
1961-01-05	87448	27.0	14.1	0.0
1961-01-06	87448	26.3	11.3	0.0
1961-01-07	87448	34.1	12.0	6.7
1961-01-08	87448	32.8	15.9	0.0
1961-01-09	87448	37.6	16.1	0.0
1961-01-10	87448	26.9	4.6	0.0

Los nombres de las variables son importantes y deben ser siempre los mismos ya que el modelo las reconocerá a partir de los mismos. Los nombres deben ser los siguientes:

- `date` : corresponde a la fecha del día en formato `Date`. El formato de la fecha para facilitar el reconocimiento por parte de R es "YYYY-MM-DD", es decir, el año expresado con cuatro dígitos y luego dos dígitos para el mes y dos para el día.
- `station_id`: Identificador unívoco de cada una de las estaciones. Debe ser un número **entero**.
- `tmax`: temperatura máxima diaria expresada en °C.
- `tmin`: temperatura mínima diaria expresada en °C.
- `prcp`: precipitación diaria expresada en mm.

El orden de las variables no es importante pero, como se mencionó, si se deben respetar los nombres de cada una. En el caso de faltantes, no se utiliza ningún valor específico para los NAs, sólo se debe dejar ese valor vacío. Este archivo tiene un formato largo, es decir, las estaciones se deben colocar una debajo de la otra.

A continuación veremos una visualización de las temperaturas máxima y mínima de la estación elegida para el ajuste.



Al tratarse de un modelo que ajusta condicionado por la variabilidad de baja frecuencia preexistente en los datos observados es necesario la agregación de las variables diarias en totales trimestrales de precipitación y medias trimestrales de temperaturas máxima y mínima. Esta operación puede realizarse con la función `summarise_seasonal_climate` incluida en el paquete. Esta función, además de agregar los datos, permite la imputación de faltantes. Se toleran una cierta cantidad que puede ser determinada por el usuario. El método de imputación utilizado es el `imputePCA()` de la librería `missMDA` .

```
# Agregación de valores diarios
seasonal_variates <- gamwgen::summarise_seasonal_climate(climate, umbral_faltantes = 0.2)

# Se muestran las primeras cinco filas
knitr::kable(seasonal_variates[1:10,])
```

station_id	year	season	seasonal_prcp	seasonal_tmax	seasonal_tmin
87448	1961	1	273.7	30.98764	14.5134831
87448	1961	2	236.0	24.47473	8.4670330
87448	1961	3	11.3	19.04565	1.3456522
87448	1961	4	234.5	24.66235	7.7611765
87448	1962	1	402.4	29.92333	14.2455556
87448	1962	2	187.2	24.30440	7.9868132
87448	1962	3	50.9	17.21957	-0.8978261
87448	1962	4	179.7	25.39560	7.8472527
87448	1963	1	306.8	29.09101	14.1404494
87448	1963	2	82.0	25.46196	8.5695652

Cabe mencionar que con esta función las valores se agregan por trimestre considerando la siguiente definición:

- Verano: Diciembre, Enero y Febrero
- Otoño: Marzo, Abril y Mayo
- Invierno: Junio, Julio y Agosto
- Primavera: Septiembre, Octubre y Noviembre

Los valores también se podrían agregar siguiendo otra definición de estaciones pero en ese caso, el usuario debería hacerlo por su cuenta. Algunas funciones útiles para hacerlo son las disponibles en el paquete `lubridate` como `quarter()` que permite definir el mes de comienzo de los trimestres. Para estas variables los nombres también son importantes por lo que deben respetarse los mostrados anteriormente.

4.2 Ajuste de los modelos estadísticos

Luego de obtener los datos observados se procede al ajuste de los modelos estadísticos: dos para temperaturas máxima y mínima y dos para precipitación. Estos modelos necesitan de datos observados diarios de variables meteorológicas. Si bien se toleran una cierta cantidad de faltantes es conveniente que las series tengan una longitud no menor a 30 años para así capturar la variabilidad climática observada.

A continuación se mostrará como ajustar los cuatro modelos estadísticos para una sola estación meteorológica condicionados por la variabilidad de baja frecuencia. La anterior es sólo uno de las configuraciones posibles y para ver más detalles de las demás posibilidades se sugiere consultar el manual completo del generador.

El ajuste del modelo local (en un punto) necesita de dos funciones: en una se define la configuración general del modelo y con la segunda se corre el modelo propiamente dicho.

Primero se crea un objeto con el control para el ajuste del simulador. Los argumentos son:

- `prcp_occurrence_threshold`: umbral de precipitación para un día lluvioso. La OMM recomienda un umbral de 0.1 mm para considerar un día como lluvioso.
- `avbl_cores`: cantidad de núcleos disponibles para la paralelización.
- `planar_crs_in_metric_coords`: sistema de coordenadas planar.

```
control_fit <-gamwgen::local_fit_control(
  prcp_occurrence_threshold = 0.1, # Umbral para la definición de días húmedos
  avbl_cores = 1, # Cantidad de núcleos disponibles
  planar_crs_in_metric_coords = 22185) # Sistema de referencia espacial (en metros)
```

Luego se corre el ajuste para la estación meteorológica con la función `local_calibrate`. Los argumentos de la función son:

- `climate`: datos meteorológicos observados para la estación

- `stations`: metadatos de las estaciones meteorológicas
- `seasonal_covariates`: datos agregados trimestrales. Si es NULL el ajuste será sin covariables y las series generadas serán estacionarias.
- `control`: objeto de control
- `verbose`: controla la impresión de mensajes en la consola. FALSE por defecto.

Nota: Si luego de leer este documento desean correr el generador con sus propios datos deben incluirlos en la carpeta `input_data` y asignarle al objeto `correr.generador` el valor TRUE. Ahora esta variable es FALSE porque los resultados del ajuste ya han sido precalculados.

```
correr.generador <- FALSE

# Al correr la función se realiza el ajuste de los cuatro modelos para cada una de
# las estaciones. En este caso, por cuestiones de tiempo a cargar un objeto ya precalculado.
# Si el usuario desea correrlo deberá ver la nota anterior.
gamgen_fit <- gamwgen::local_calibrate(climate = climate, # Registro histórico de variables meteorológi
  stations = stations, # Estaciones meteorológicas
  seasonal_covariates = NULL, # Totales trimestrales de precipitación
  control = control_fit, # Objeto de control
  verbose = FALSE) # Impresión de mensajes en la consola.
```

Para esta demostración, cargamos el objeto con el ajuste del modelo ya realizado.

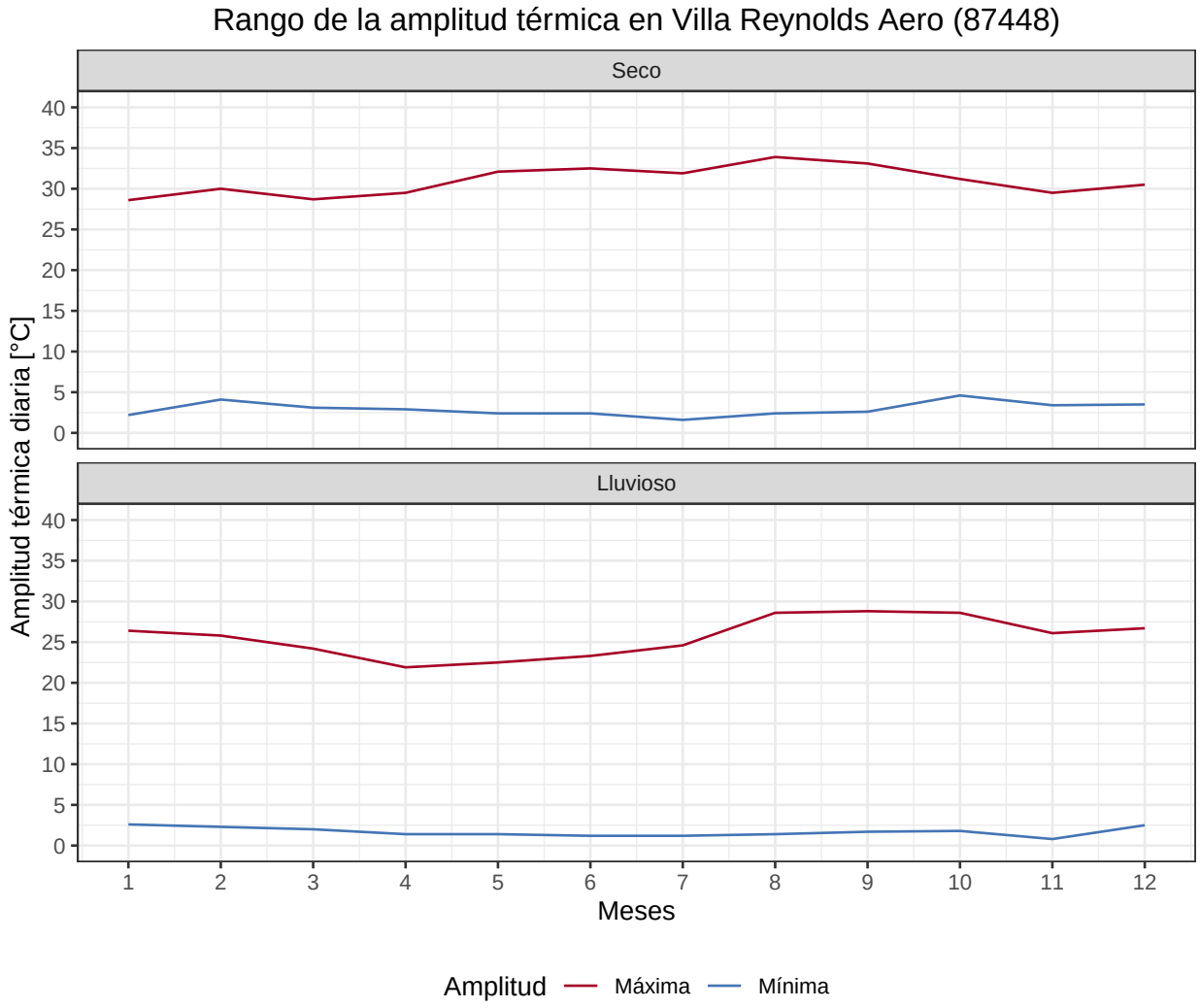
```
# Copiamos el archivo preajustado a nuestro directorio de trabajo
if (!fs::file_exists('input_data/local/fit_local.RData')) {
  fs::file_copy(system.file('/autorun/local', "fit_local.RData",  package = "gamwgen"),
                new_path = 'input_data/local/fit_local.RData')
}

# Cargamos el archivo recientemente creado
load('input_data/local/fit_local.RData')

# Clase del objeto con el ajuste del generador
class(gamgen_fit)

## [1] "gamwgen"

# Contenido del modelo
names(gamgen_fit)

##  [1] "control"              "stations"             "climate"
##  [4] "seasonal_covariates"  "crs_used_to_fit"      "start_climatology"
##  [7] "fitted_models"        "models_data"          "models_residuals"
## [10] "statistics_threshold" "exec_times"
```

Dentro del objeto se guardan todo lo necesario para la simulación así como información accesoria.

- `control`: copia de la configuración usada para calibrar el generador
- `stations`: estaciones meteorológicas utilizadas para la calibración
- `climate`: datos climáticos de cada uno de las estaciones
- `seasonal_covariates`: series temporales de totales trimestrales de precipitación y medias trimestrales de temperaturas máxima y mínima.

- `crs_used_to_fit`: sistema de referencia espacial usado para proyectar
- `start_climatology`: climatología diaria de cada una de las variables de entrada.
- `fitted_models`: modelos ajustados, uno para cada variable: temperaturas máxima y mínima y ocurrencia y montos de precipitación.
- `models_data`: datos usados efectivamente usados para ajustar los modelos (sin NAs)
- `models_residuals`: residuos de cada uno de los modelos. Es decir, la diferencia entre el valor ajustado por el modelo (clima local) y el valor observado en el día **i**
- `statistics_threshold`: umbrales de amplitud térmica diaria por mes. Si la amplitud simulada está fuera de este rango, se repetirá la simulación para ese día a los fines de mantener la consistencia entre variables
- `exec_times`: tiempo de ejecución de cada una de las etapas del ajuste

Los umbrales de amplitudes máximos y mínimos permitidas son muy importantes para evitar que se produzcan temperaturas máximas inferiores a las mínimas. En la siguiente figura se puede ver la variabilidad mensual de cada uno de estos umbrales y el efecto del tipo de día (seco o lluvioso) sobre ellos.



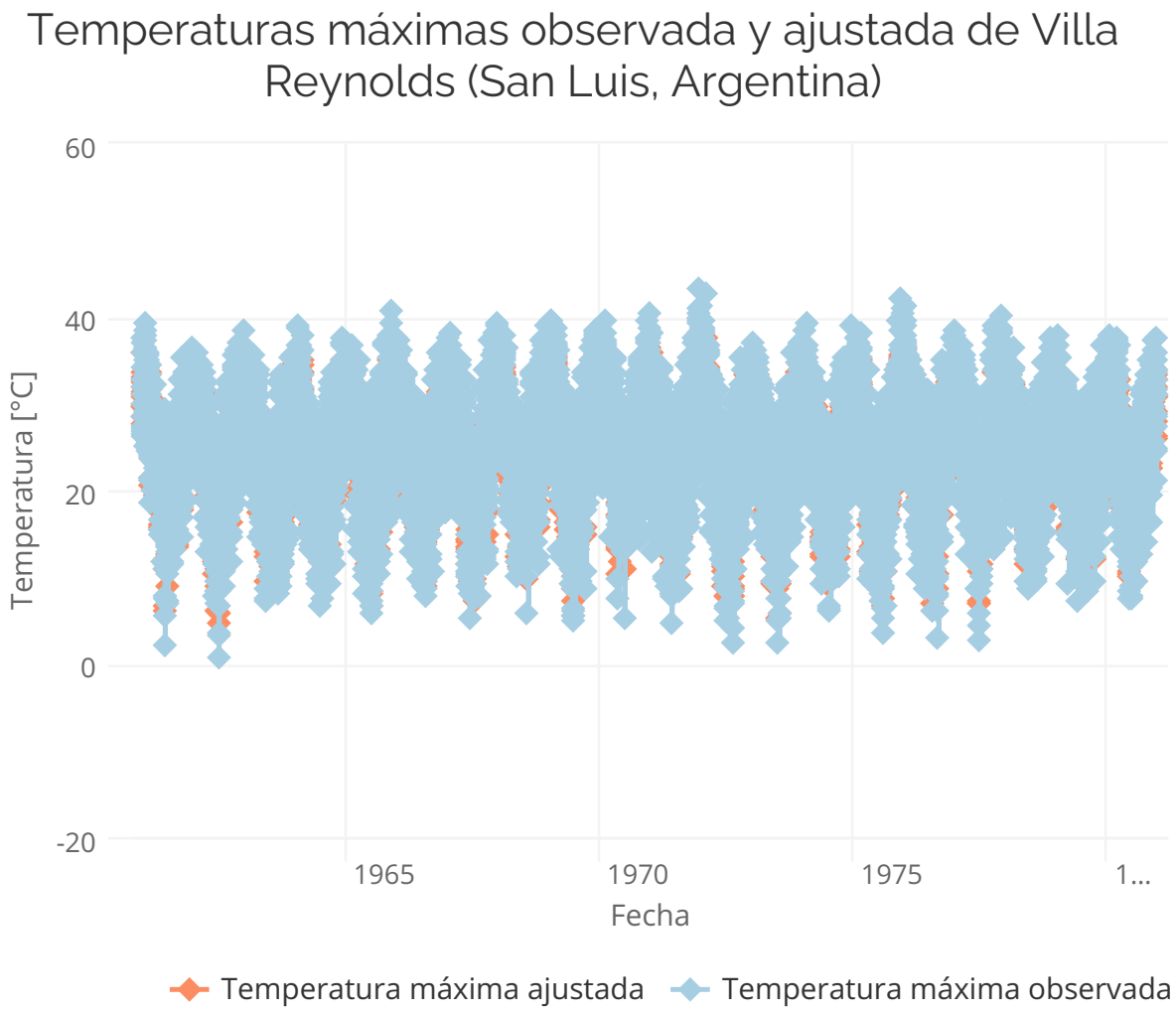
Cada uno de los GAMs ajustados se almacenan en el objeto `gamgen_fit` y pueden ser evaluados con la función `summary()`.

```r
summary(gamgen_fit$fitted_models$`87448`$tmax_fit)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## tmax ~ s(tmax_prev, tmin_prev, k = 50) + s(prcp_occ, bs = "re") +
##     s(prcp_occ_prev, bs = "re") + s(doy, bs = "cc", k = 30) +
##     s(SX1, SN1, k = 20) + s(SX2, SN2, k = 20) + s(SX3, SN3, k = 20) +
##     s(SX4, SN4, k = 20)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.62116    0.03189   803.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df       F p-value
## s(tmax_prev,tmin_prev) 29.9069 38.405  225.74  <2e-16 ***
## s(prcp_occ)             0.9989  1.000  930.30  <2e-16 ***
## s(prcp_occ_prev)        0.9995  1.000 2248.11  <2e-16 ***
## s(doy)                 12.6336 28.000   73.12  <2e-16 ***
## s(SX1,SN1)              3.0997  3.689   55.32  <2e-16 ***
## s(SX2,SN2)              2.0001  2.000   89.25  <2e-16 ***
## s(SX3,SN3)              4.4053  5.853   35.79  <2e-16 ***
## s(SX4,SN4)              2.0001  2.000   86.85  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.713   Deviance explained = 71.4%
## fREML =  58829  Scale est. = 13.964    n = 21466
```

La función `summary` permite analizar los resultados del ajuste de cada uno de los modelos. Para el caso del modelo de temperatura máxima podemos ver la fórmula del GAM en la parte superior bajo el apartado `Formula` y la significancia de cada uno de los términos del modelo en la tabla inmediatamente inferior. Se puede observar que todos los términos son altamente significativos. También se incluyen como pruebas de bondad del ajuste el porcentaje de la varianza explicada por el modelo y el valor de R-ajustado.

Ahora vamos a visualizar los resultados del modelo ajustado de temperatura máxima (clima local) y la temperatura observada de la estación.



Recordemos que las diferencias entre ambas curvas, los residuos del modelo, nos permitirán modelar el tiempo local que era la componente estocástica del generador. Una vez que el los distintos modelos estadísticos de temperaturas y precipitación estén ajustados podemos proseguir a la generación de las series sintéticas.

5 Generación de series sintéticas

La simulación de datos también está compuesta por dos funciones. En una se especifica gran parte de la configuración de la simulación mientras que la otra realiza la simulación propiamente dicha.

Los argumentos de la función de control son:

- `nsim`: cantidad de simulaciones a realizar. Se debe ingresar un valor **entero** mayor o igual a 1.
- `seed`: semilla. Se debe ingresar cualquier numero **entero**. No es necesario recordarlo porque se guarda junto a los resultados.
- `avbl_cores`: cantidad de núcleos disponibles para la paralelización.
- `use_spatially_correlated_noise`: utilizar la generación estocástica espacialmente correlacionada. Esta opción sólo es válida si en el ajuste y en la simulación se usaron más de cinco estaciones meteorológicas diferentes. Con un menor número no es posible calcular los variogramas necesarios para la generación de los campos aleatorios. Se debe introducir un **boolean** (TRUE or FALSE).

- `use_temporary_files_to_save_ram`: si se simulan muchas realizaciones o los recursos informáticos son escasos, esta opción permite guardar los resultados de cada una de las realizaciones en el disco liberando memoria RAM que quedará disponible para generar nuevas simulaciones. Al finalizar la generación todos los archivos se combinan en uno único. Se debe introducir un **boolean** (TRUE or FALSE).
- `use_temporary_files_to_save_ram`: esta opción permite eliminar los archivos temporales creados para ahorrar RAM luego de terminar la generación de todas las simulaciones. Se debe introducir un **boolean** (TRUE or FALSE).

```
control_sim <- gamwgen::local_simulation_control(
  nsim = 100, # Cantidad de simulaciones a realizar
  seed = 1234, # Semilla para que los resultados sean reproducibles
  avbl_cores = 1, # Cantidad de núcleos disponibles a utilizar
  use_spatially_correlated_noise = FALSE, # Usar modelo de ruido espacialmente correlacionado
  use_temporary_files_to_save_ram = FALSE, # Guardar resultados intermedios para ahorrar RAM
  remove_temp_files_used_to_save_ram = TRUE) # Borrar los resultados intermedios creados anteriormente
```

Luego se procede a la simulación de datos meteorológicos. Los argumentos de la función de simulación son:

- `model`: objeto con el resultado de la función `local_calibrate()`
- `simulation_locations`: objeto tipo `sf` con la ubicación de las estaciones a simular. Las estaciones usadas deben haber sido incluidas en el proceso de ajuste. No es necesario que todas estén presentes, se pueden generar series solo sobre algunas de ellas.
- `start_date`: fecha de comienzo de la generación de series sintéticas. Si no se incluyeron covariables estacionales en el ajuste, la fecha de comienzo es completamente arbitraria. Caso contrario, la fecha de comienzo no puede ser anterior al inicio de la serie de covariables, ni tampoco posterior. Se debe introducir una fecha en formato **date**
- `end_date`: fecha de fin de la generación de series sintéticas. Si no se incluyeron covariables estacionales en el ajuste, la fecha de fin es completamente arbitraria. Caso contrario, la fecha de comienzo no puede ser anterior al inicio de la serie de covariables, tampoco puede ser posterior. Se debe introducir una fecha en formato **date**
- `control`: objeto de control creado con la función `control_sim()`.
- `output_folder`: ruta al directorio donde se guardarán los resultados, tanto finales como intermedios.
- `output_filename`: nombre del archivo de salida. Para facilitar la interoperabilidad, el archivo generado es un archivo de texto en formato separado por comas (.csv)
- `seasonal_covariates`: datos agregados trimestrales. Si el ajuste se realizó con covariables, la generación también debe realizarse con ellas. Caso contrario se producirá un error. Se debe introducir un data frame con los valores agregados para las tres variables (precipitación y temperaturas máxima y mínima) pero no necesariamente deben ser los mismos a los utilizados en el ajuste. Si se desean simular tendencias de algún tipo, ya sea de un modelo de cambio climático o arbitrarias, se deben perturbar estas variables trimestrales e introducirlas aquí. Estas series si deben tener la misma longitud que el período a generar
- `verbose`: controla la impresión de mensajes en la consola. FALSE por defecto.

Nota: Si luego de leer este documento desean correr el generador con sus propios datos deben incluirlos en la carpeta `input_data` y asignarle al objeto `correr.generador` el valor TRUE. Ahora esta variable es FALSE porque los resultados del ajuste ya han sido precalculados. También debe especificarse la ruta a la carpeta `ouput_data` para así guardar los resultados y poder correr la validación.

```
correr.generador <- FALSE

# Al correr la función se realiza la generación de series para cada una de las estaciones.
# En este caso, por cuestiones de tiempo, vamos a cargar un objeto con los resultados de la simulación
```

```r
simulated_climate <- gamwgen::local_simulation(model = gamgen_fit, # Objeto con los resultados del ajus
    simulation_locations = stations, # Estaciones para las cuales simular
    start_date = as.Date('1961-01-01'), # Fecha de comienzo de las simulaciones
    end_date = as.Date('2019-01-01'), # Fecha de fin de las simulaciones
    control = control_sim, # Objeto con la configuración
    output_folder = getwd(), # Directorio donde se guardarán los resultados
    output_filename = 'simulations.csv', # Nombre del archivo de salida
    seasonal_covariates = seasonal_covariates, # Covariables estacionales
    verbose = FALSE) # Impresión de mensajes en la consola
```

Esta función produce dos tipos de resultados: una lista que permanece en el ambiente de R y los datos generados que son guardados como .csv en el directorio indicado precedentemente.

```r
# Copiamos el archivo preajustado a nuestro directorio de trabajo
if (!fs::file_exists('output_data/local/simulated_climate_local.RData')) {
  fs::file_copy(system.file('/autorun/local', "simulated_climate_local.RData",  package = "gamwgen"),
              new_path = 'output_data/local/simulated_climate_local.RData')
}
# Cargamos el archivo recientemente creado
load('output_data/local/simulated_climate_local.RData')

# Clase del objeto con el ajuste del generador
class(simulated_climate)
```

```
## [1] "list"            "gamwgen.climate"
```

```r
# Contenido del modelo
names(simulated_climate)
```

```
## [1] "nsim"
## [2] "seed"
## [3] "realizations_seeds"
## [4] "simulation_points"
## [5] "output_file_with_results"
## [6] "output_file_fomart"
## [7] "rdata_file_with_fitted_stations_and_climate"
## [8] "exec_times"
```

La lista contiene los siguientes objetos:

- `nsim`: cantidad de realizaciones.
- `seed`: semilla general para toda la generación. Corresponde a la que se incluye en la función de control.
- `realization_seeds`: semillas para cada una de las realizaciones. Esto permite replicar los resultados.
- `simulation_points`: puntos donde se generaron las series sintéticas.
- `output_fil_with_results`: nombre del archivo con los resultados.
- `output_file_format`: tipo de archivo de salida, en este caso .csv.
- `rdata_file_with_fitted_stations_and_climate`: archivo `.RData` con los datos meteorológicos observados que fueron utilizados en el ajuste. También se incluyen los metadatos de cada uno de esos puntos.
- `exec_times`: tiempo de ejecución del la generación.

Ahora veremos el formato del archivo de salida que contiene las series sintéticas.

5.1 Resultado de la generación

```
# Se carga el set de datos simulados
simulated_climate <- readr::read_csv(here::here('output_data/local/simulated_local_conditional.csv'))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   realization = col_double(),
##   station_id = col_double(),
##   point_id = col_double(),
##   longitude = col_double(),
##   latitude = col_double(),
##   date = col_date(format = ""),
##   tmax = col_double(),
##   tmin = col_double(),
##   prcp = col_double()
## )
```

```
# Primeras filas del objeto de salidas
knitr::kable(simulated_climate[1:10,])
```

realization	station_id	point_id	longitude	latitude	date	tmax	tmin	prcp
1	87448	1	5001636	6256577	2010-01-01	28.05965	17.69135	7.534622
1	87448	1	5001636	6256577	2010-01-02	27.32783	16.20019	0.000000
1	87448	1	5001636	6256577	2010-01-03	35.01025	13.35766	0.000000
1	87448	1	5001636	6256577	2010-01-04	34.18990	18.04710	0.000000
1	87448	1	5001636	6256577	2010-01-05	30.59661	20.62692	0.000000
1	87448	1	5001636	6256577	2010-01-06	37.15162	17.44571	0.000000
1	87448	1	5001636	6256577	2010-01-07	35.65166	18.36680	0.000000
1	87448	1	5001636	6256577	2010-01-08	28.28271	13.77686	0.000000
1	87448	1	5001636	6256577	2010-01-09	29.66045	13.73328	0.000000
1	87448	1	5001636	6256577	2010-01-10	28.07379	14.17866	11.043809

El resultado de la generación es un archivo `.csv` que contiene la siguiente información:

- `realization`: número de realización. Es un valor entero entre 1 y la cantidad de realizaciones definida por el usuario.
- `station_id`: número unívoco de identificación de la estación meteorológica o del punto arbitrario.
- `date`: fechas de cada uno de los días de la simulación.
- `tmax`: valores de temperatura máxima generada expresada en °C.
- `tmin`: valores de temperatura mínima generada expresada en °C.
- `prcp`: valores de precipitación diaria generada expresada en mm.

6 Diagnósticos

A continuación se cargan las funciones necesarias para la elaboración de los diagnósticos y se definen las variables que serán evaluadas.

```
# Funciones para realizar los diagnósticos
source("./src/funciones_validacion.R", local = knitr::knit_global())
```

```r
# Se definen las variables a validar
variables <- c('tmax', 'tmin', 'prcp')

# Se agrega la latitud como una variable más en el data frame con los metadatos de las estaciones
stations <- stations %>%
    dplyr::mutate(lat_dec = sf::st_coordinates(geometry)[,'Y'])
# Estaciones usadas en el ajuste. En este ejemplo es la misma porque solo se ajustó una estación
fit_stations <- stations
# Se debe definir el umbral de lluvia
umbral.precipitacion <- 0.1
```

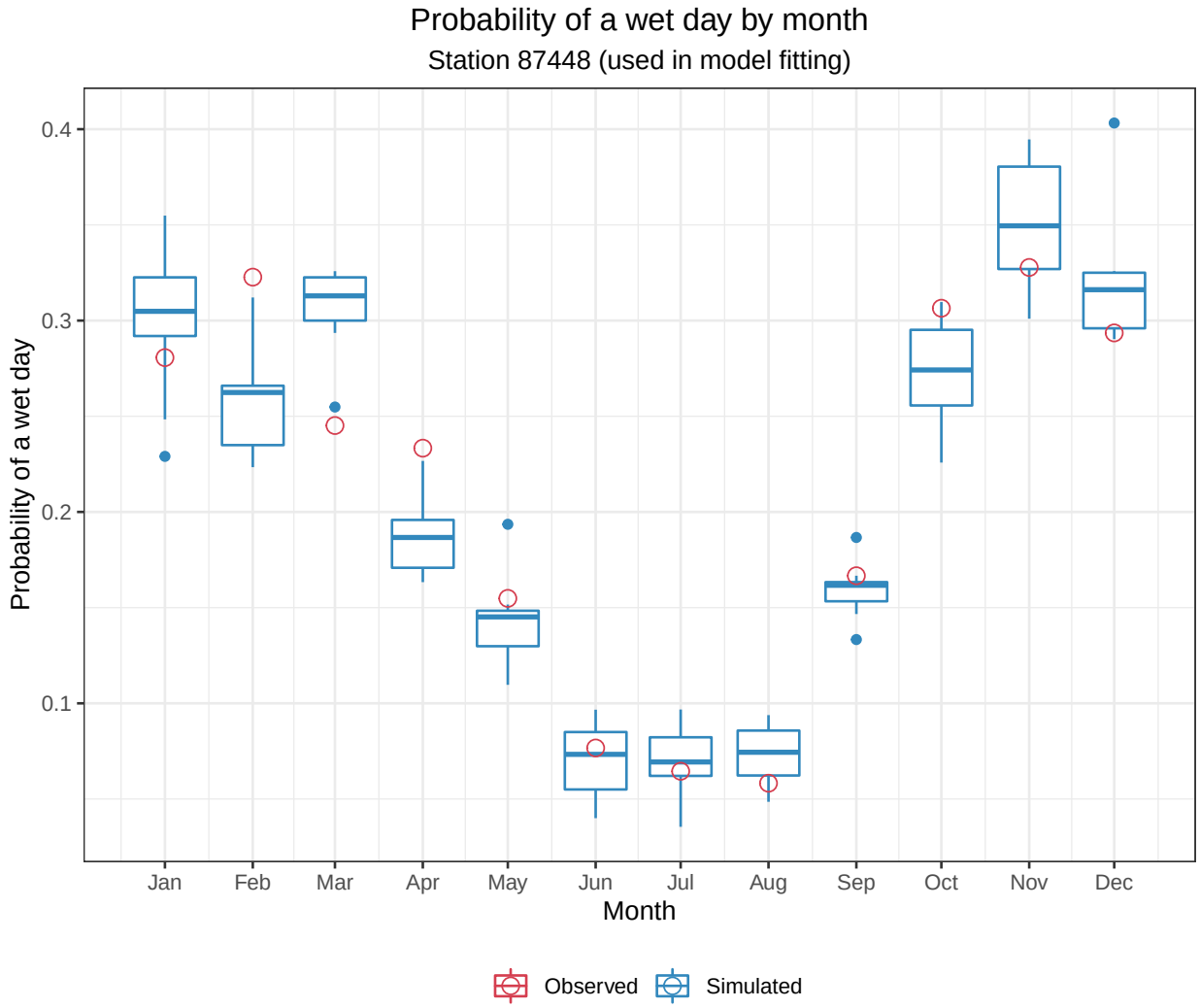
El objetivo de las distintas pruebas diagnósticas es comprobar si los datos observados pueden ser considerados una realización más del generador. En otras palabras, que las series observadas no pueden distinguirse de las sintéticas. Esto no implica que tengan exactamente la misma media o desvío estándar, etc.

```
## Estacion 87448
```

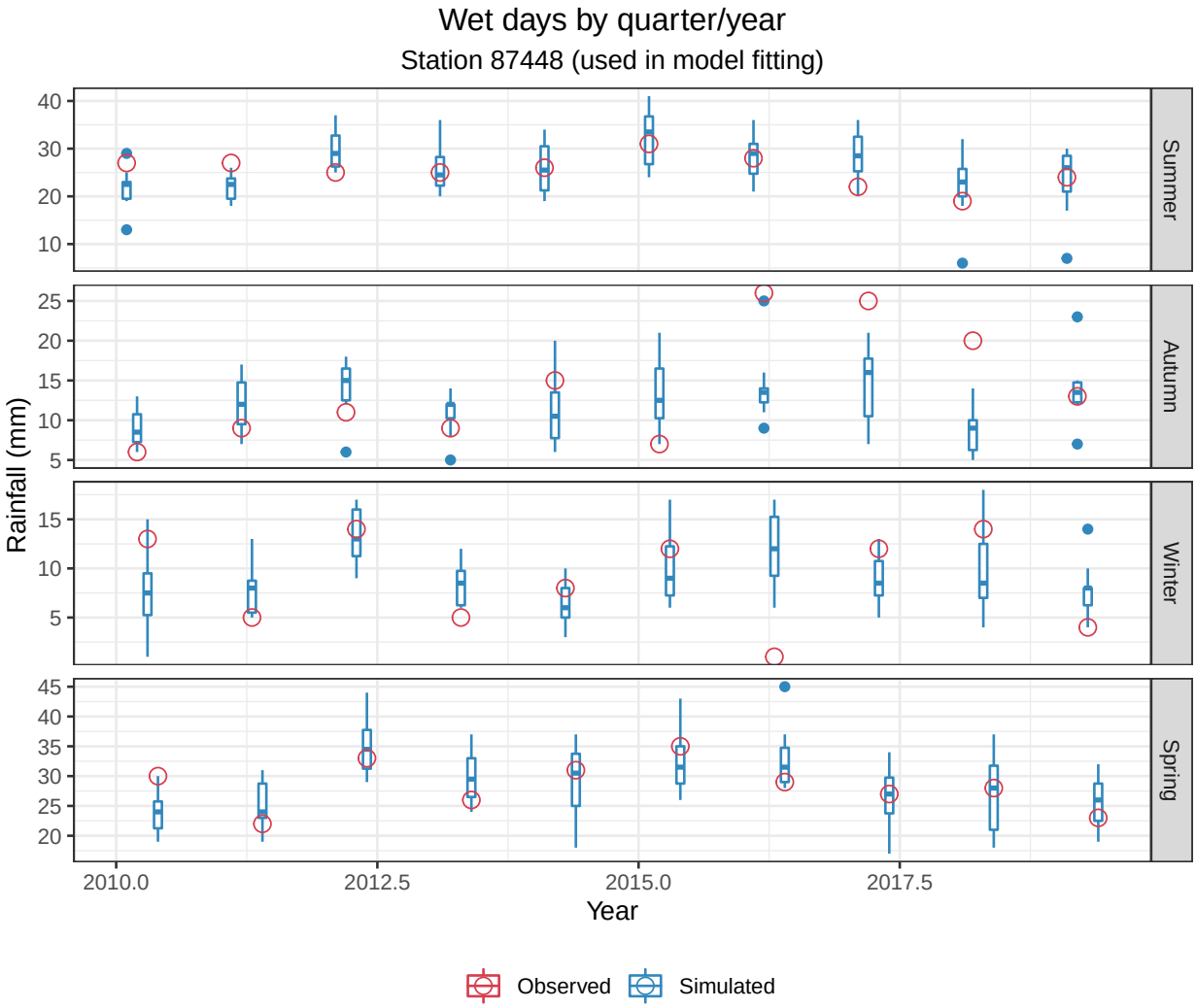
```
## Procesando estacion 87448
## ... variable tmax
## ... variable tmin
## ... variable prcp
## ... other plots
```

6.1 Diagnósticos de precipitación

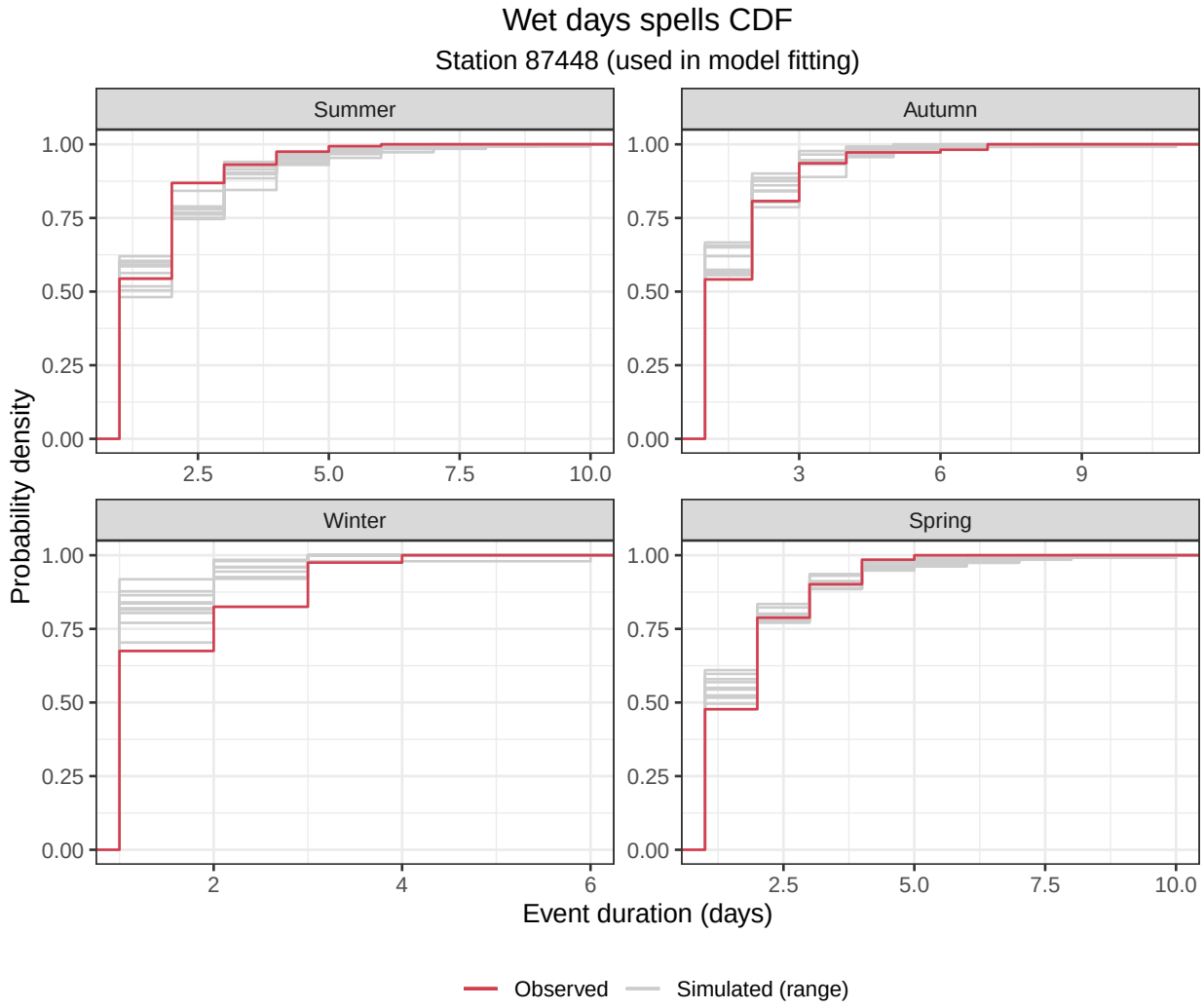
A continuación se mostrarán los distintos diagnósticos desarrollados para validar las series diarias de precipitación.



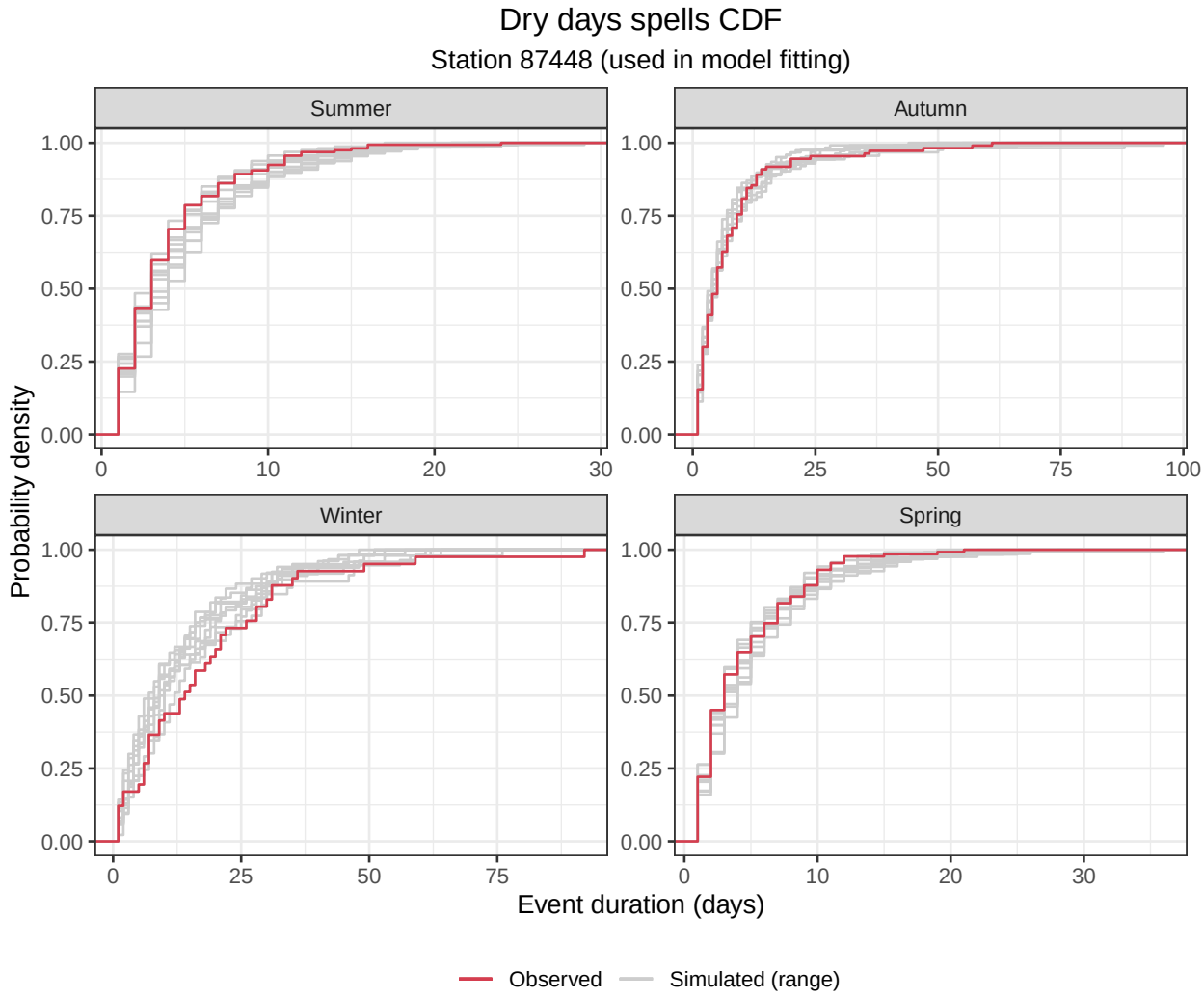
Este diagrama de cajas muestra la probabilidad de ocurrencia de un día húmedo o lluvioso por mes. Las cajas corresponden a las distintas realizaciones y los puntos rojos corresponden al valor observado en la serie histórica. Se observa un claro patrón estacional y como el generador es capaz de capturarlo. En verano, uno de cada tres días es lluvioso mientras que en invierno la probabilidad baja a menos de uno en diez.



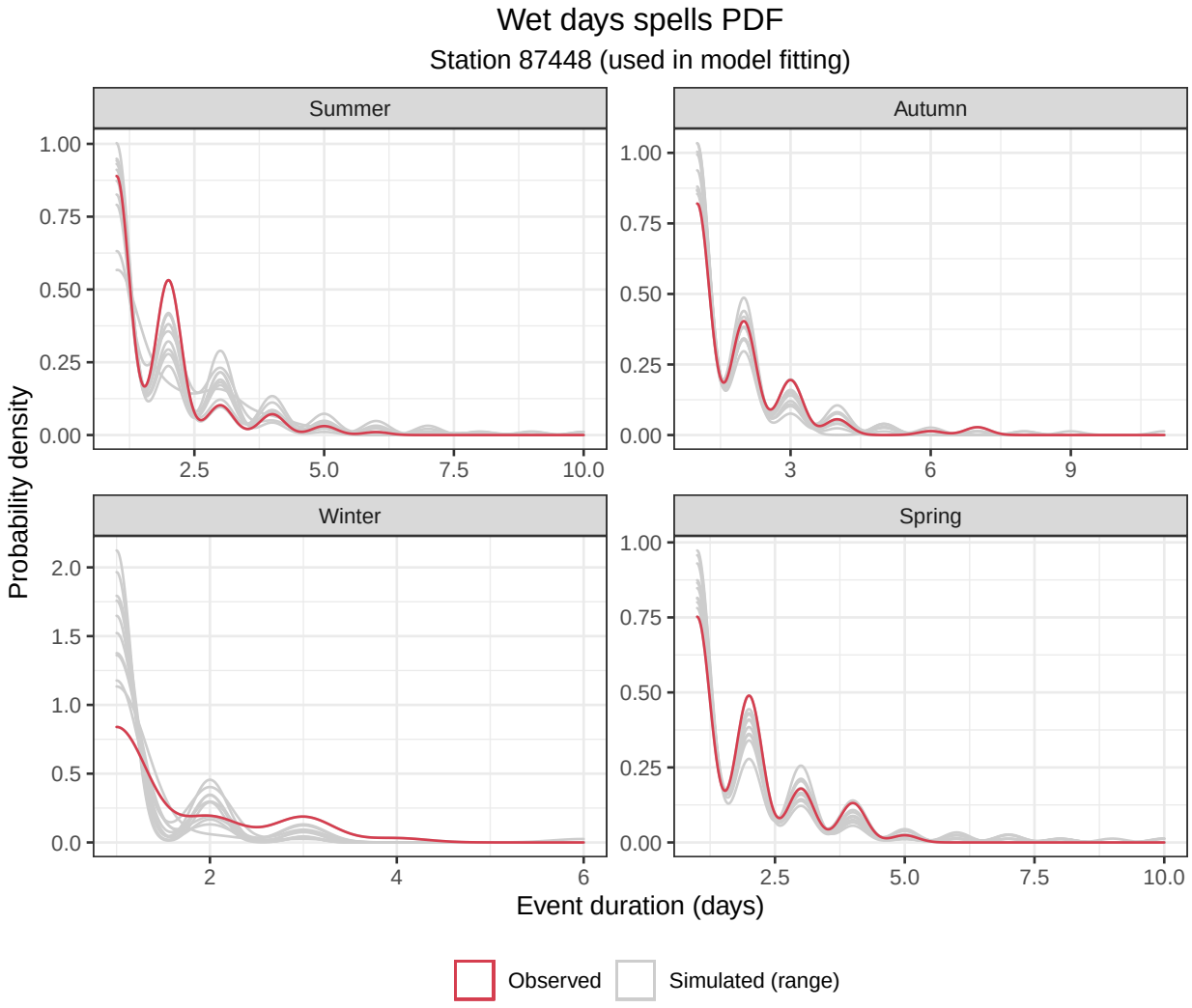
Esta figura muestra la cantidad de días lluviosos por trimestre a lo largo del tiempo. Los puntos rojos corresponden a la cantidad de días lluviosos observados mientras que las cajas corresponden a las distintas realizaciones. Se observa que los puntos se encuentran dentro del rango de las distintas cajas a excepción de un trimestre excepcionalmente húmedo a principios de los 2000.



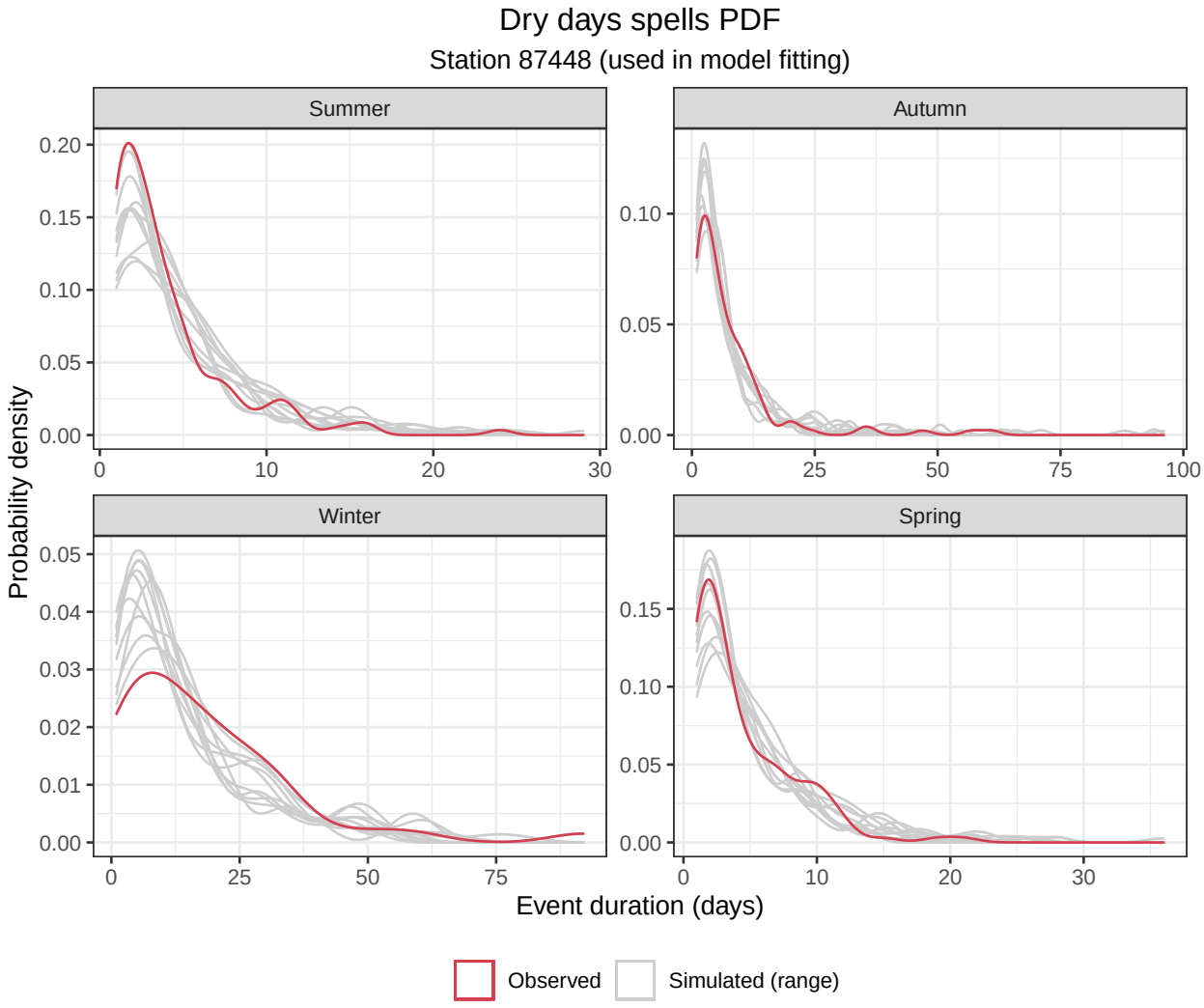
Esta figura muestra la probabilidad acumulada de ocurrencia de rachas secas por trimestre. La línea roja corresponde a la probabilidad acumulada observada mientras que la envolvente gris corresponde a cada una de las realizaciones. Es un indicador muy importante para el análisis del riesgo de sequía ya que índica si la duración de los déficits hídricos es capturada o no.



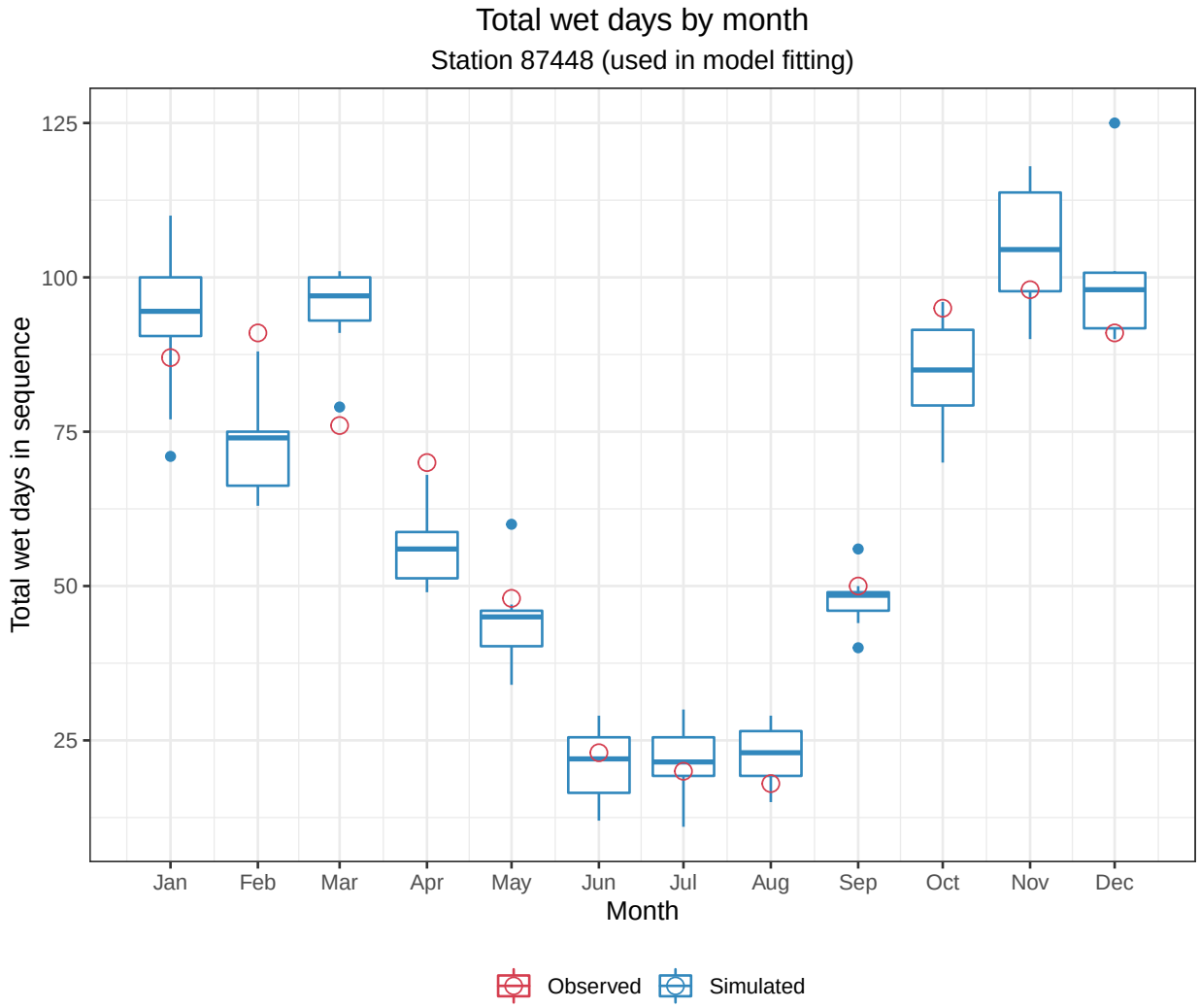
Esta figura es muy similar a la anterior con la diferencia que considera las rachas secas en lugar de las secas. La línea roja corresponde a los datos observados y la envolvente gris a las distintas realizaciones. Se observa que para los cuatro trimestres la línea roja se pierde entre las realizaciones.



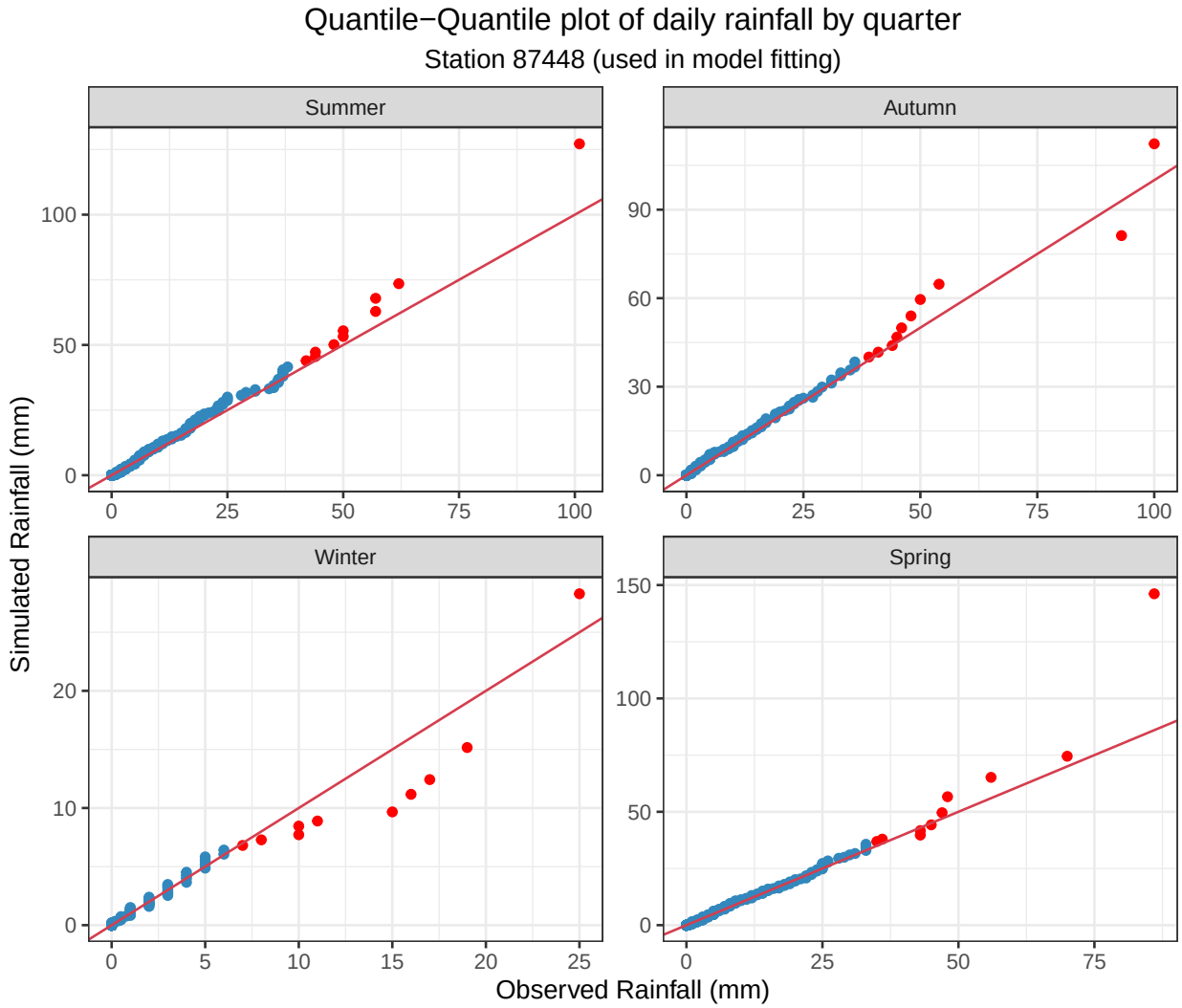
Esta figura muestra la probabilidad de ocurrencia de rachas lluviosas de una determinada longitud de días por trimestre. En otras palabras, permite saber cuán frecuente es una racha lluviosa en esta estación meteorológica para cada trimestre. Las líneas rojas corresponden a la serie observada mientras que la envolvente gris a las distintas realizaciones.



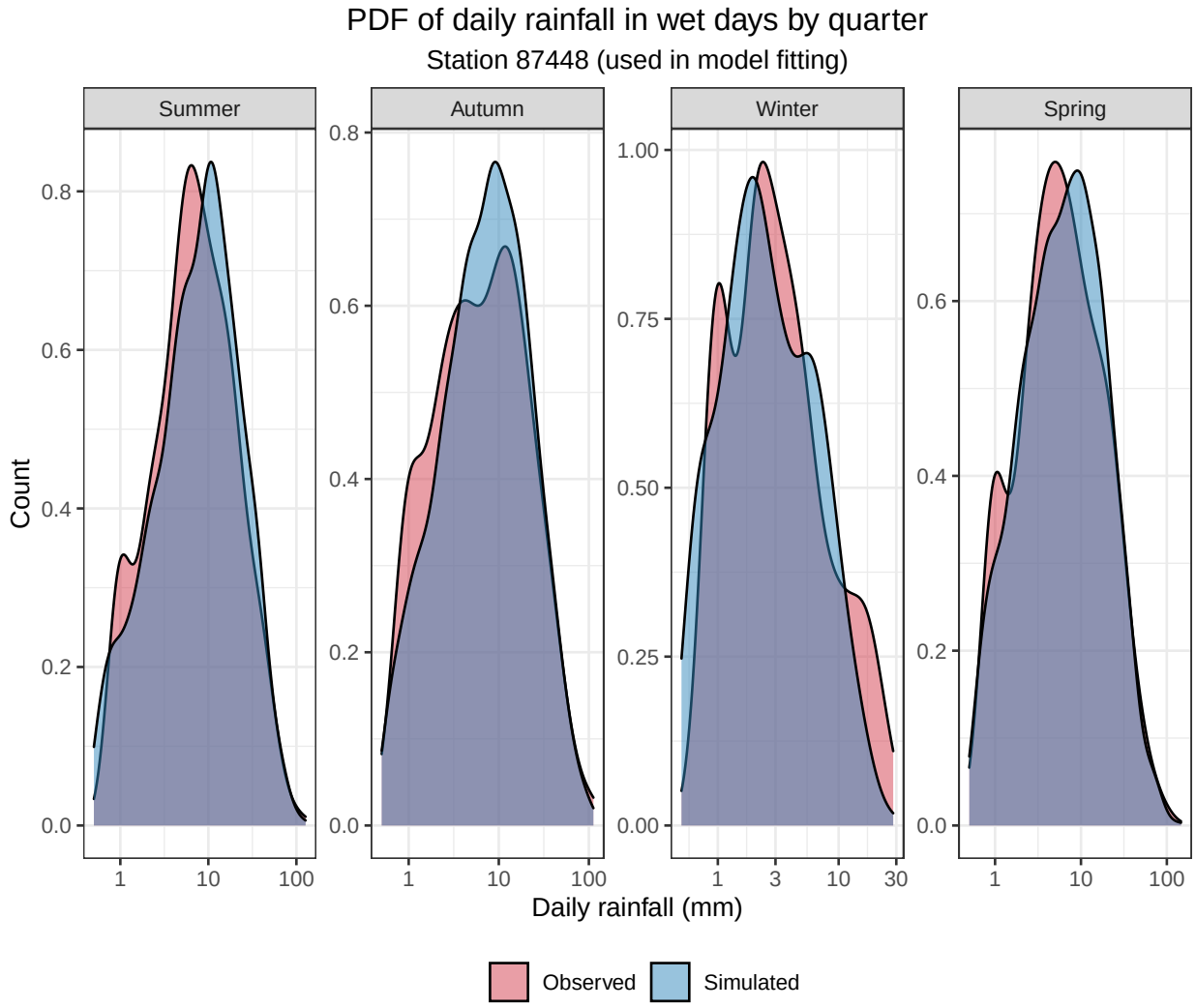
Esta figura es análoga a la anterior solo que considera las rachas secas. Las líneas rojas corresponden a la serie observada y la envolvente gris a las distintas realizaciones. Se observa que las realizaciones tienen el mismo comportamiento que las series observadas y que se captura de manera correcta la variación en las rachas para los distintos trimestres. Para estudios de impacto de sequía este diagnóstico es particularmente importante ya que será uno de los determinantes principales de los déficit acumulados.



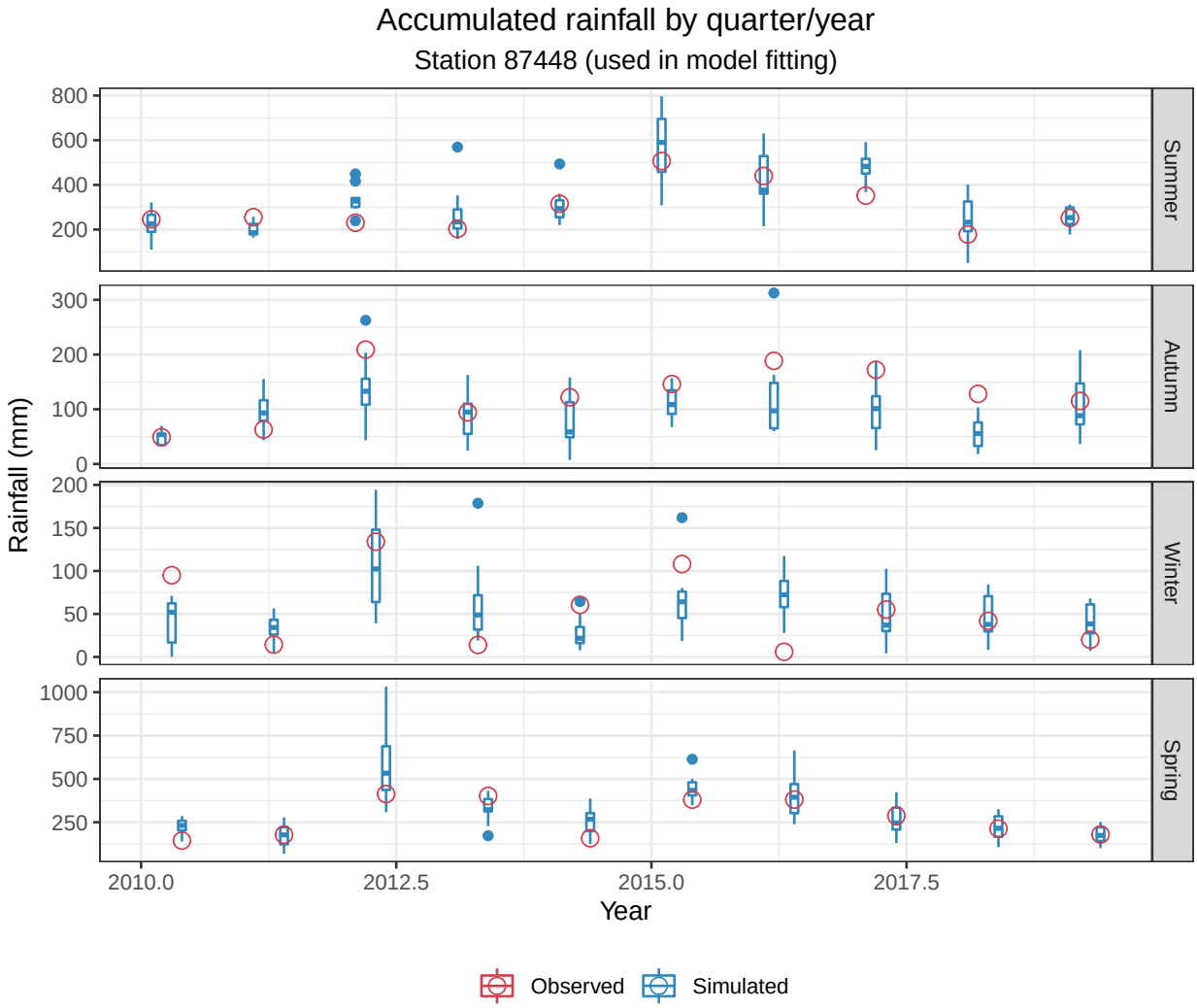
Este diagrama de cajas muestra la cantidad de días lluviosos por mes y su variabilidad a lo largo del año. Las cajas corresponden a las distintas realizaciones mientras que los puntos rojos a los valores observados. La cantidad de días lluviosos por mes se acumula para todos los años del período, en este caso 57 años, es por esto que en enero los días lluviosos totalizaron aproximadamente 550. A pesar de existir una diferencia de casi 400 días entre los meses estivales e invernales, el generador es capaz de capturar están gran diferencia.



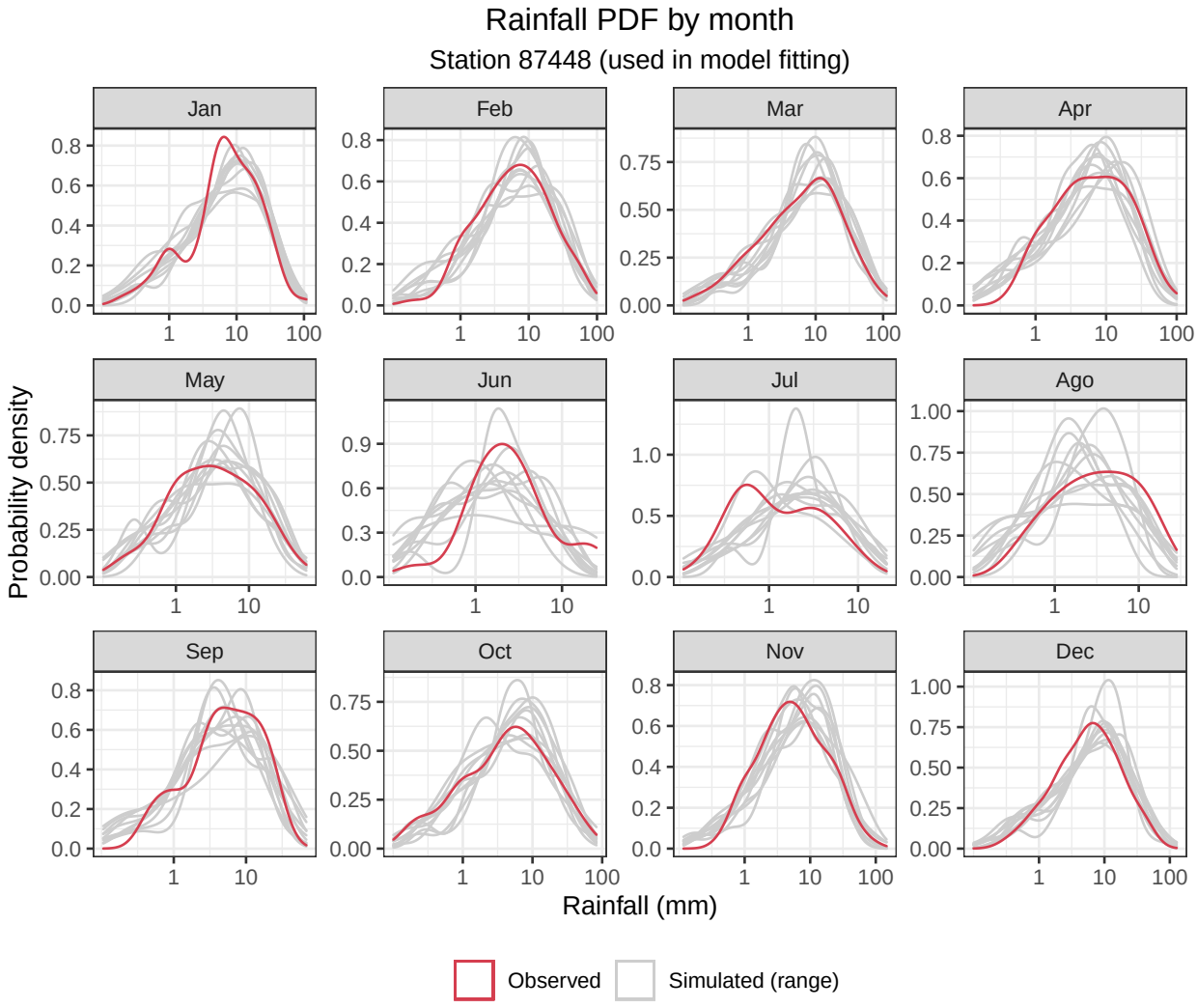
Esta figura compara los cuantiles observados de precipitación diaria y los simulados por el generador por cada trimestre del año. Los puntos azules corresponden a los montos menores al percentil 99 mientras que los rojos corresponden a los montos mayores al percentil 99, es decir, a los extremos. En el eje x se ubican los cuantiles observados mientras que en el y, los simulados. El objetivo de esta prueba es que los puntos estén alineados sobre la recta 1:1 lo que indicaría que ambos cuantiles, observados y simulados, son iguales. Se observa que para casi todos los puntos, a excepción de los más extremos, la concordancia de ambas series es muy buena.



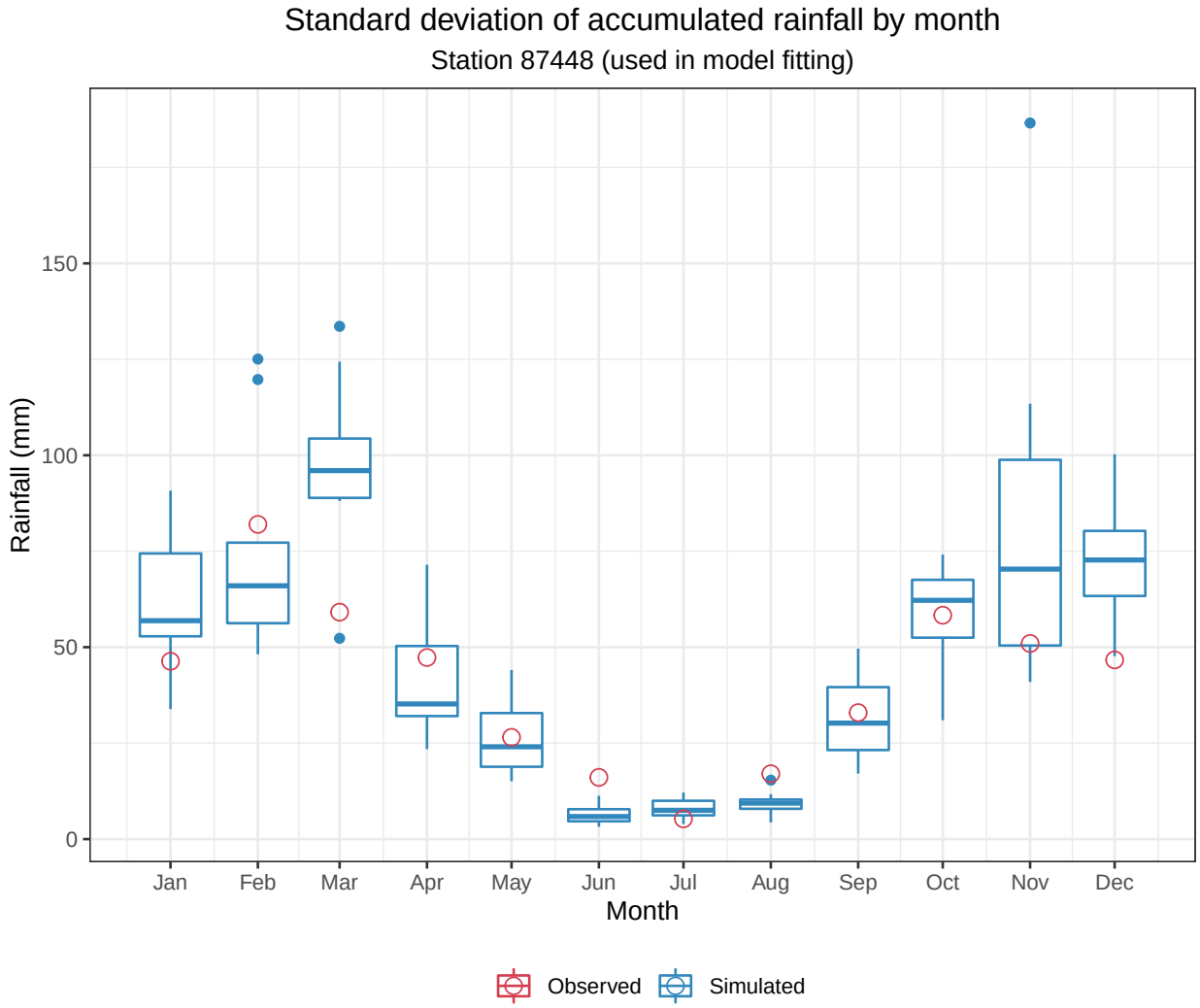
Esta figura muestra la probabilidad de los montos diarios de precipitación diaria por trimestre. El área roja corresponde a los montos observados mientras que la azul, a los simulados. Cabe mencionar que la escala en el eje x es logarítmica por lo que pequeños desplazamientos sobre el eje implican grandes cambios en los montos. Se observa un buen acuerdo para las cuatro estaciones del año a excepción de los pequeños montos cercanos a un 1 mm.



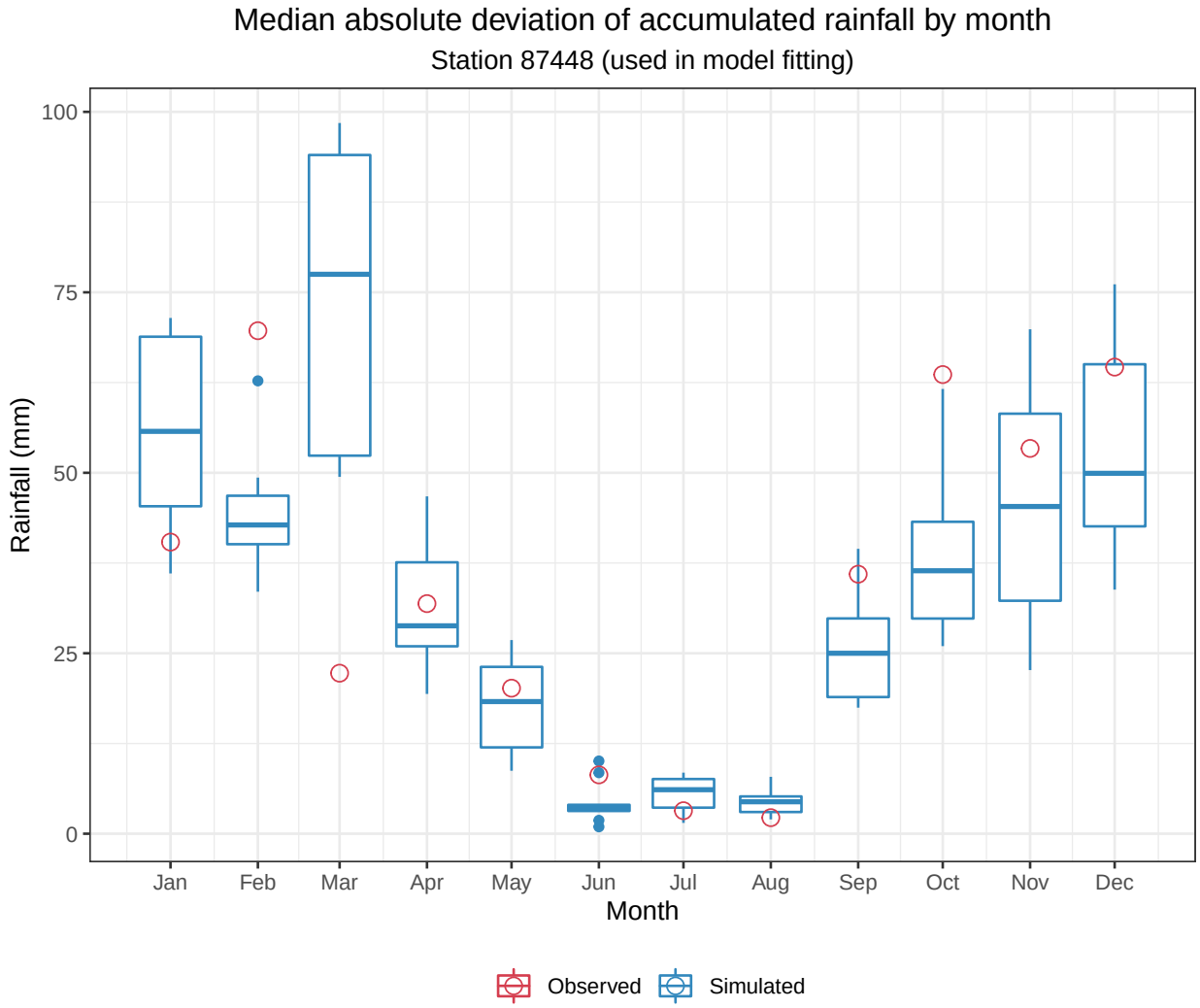
Este diagnóstico compara la precipitación agregada por trimestre. Las cajas corresponden a las distintas realizaciones mientras que los puntos rojos corresponden al valor observado. Al utilizar totales trimestrales en el ajuste y la generación se producen series que copian las variaciones observadas en el registro histórico. Es decir, si el año observado tuvo precipitaciones por debajo del promedio, las distintas realizaciones tenderán a ser menores al promedio y viceversa.



Esta figura es similar a la anterior y muestra la distribución de la precipitación por mes de precipitación. La línea roja corresponde a los montos observados mientras que las grises a las distintas realizaciones. Cabe mencionar que la escala del eje x es logarítmica por lo que pequeños desplazamientos sobre el eje implican grandes cambios en los montos. Se observa un buen acuerdo para todos los meses a excepción de montos muy bajos cercanos a 1 mm.



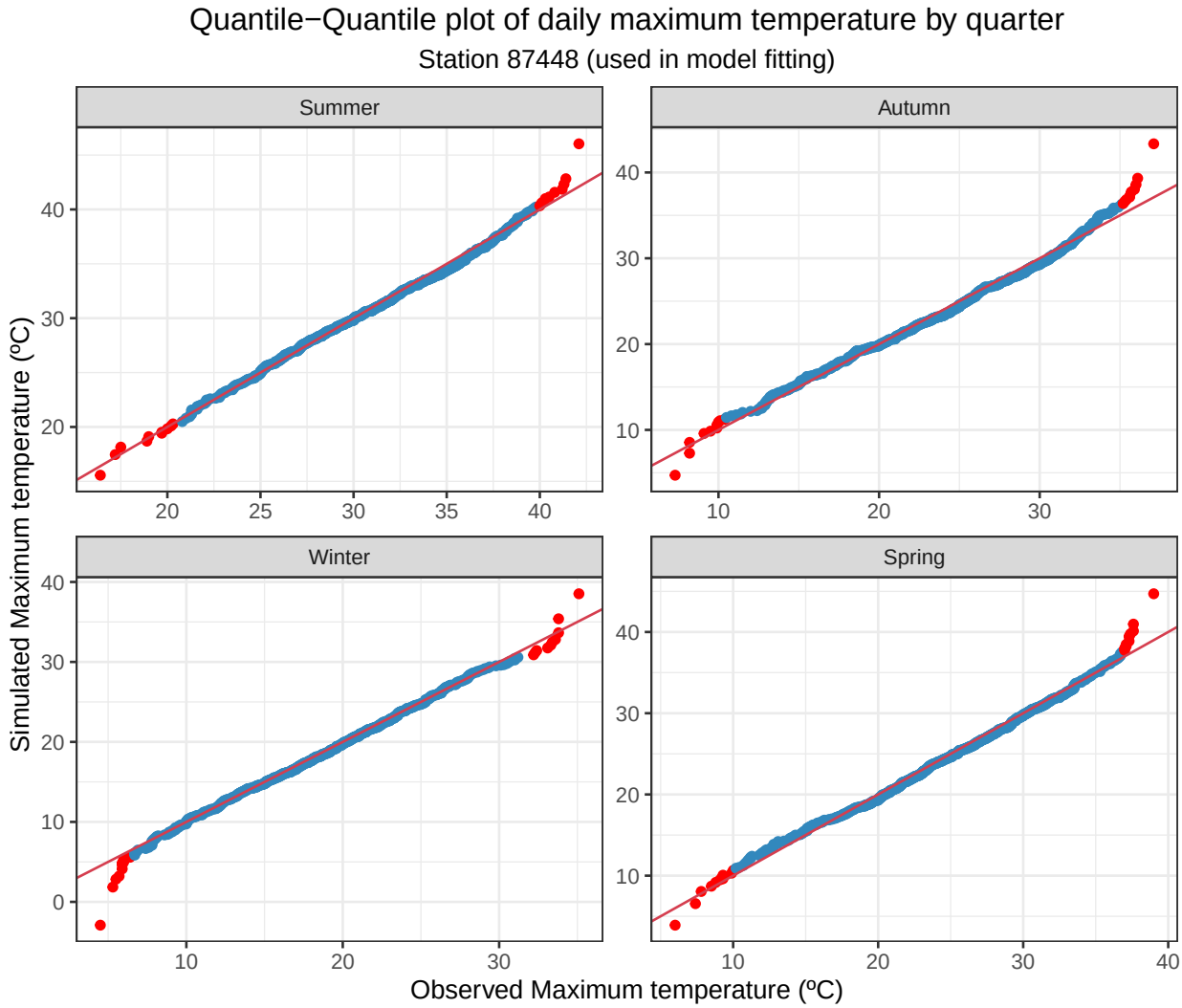
Esta figura muestra la marcha mensual de los desvíos estándar de precipitación mensual. Es decir, la variabilidad de los acumulados mensuales a lo largo del año. Los puntos rojos corresponden a los valores calculados a partir de los datos observados y las cajas a las distintas realizaciones. Se observa que el generador es capaz de capturar la variabilidad intranual ya que la mediana de las cajas siguen el mismo patrón que los datos observados. Además, las cajas en los meses estivales son mas grandes lo que implica una mayor variabilidad en comparación con los meses invernales.



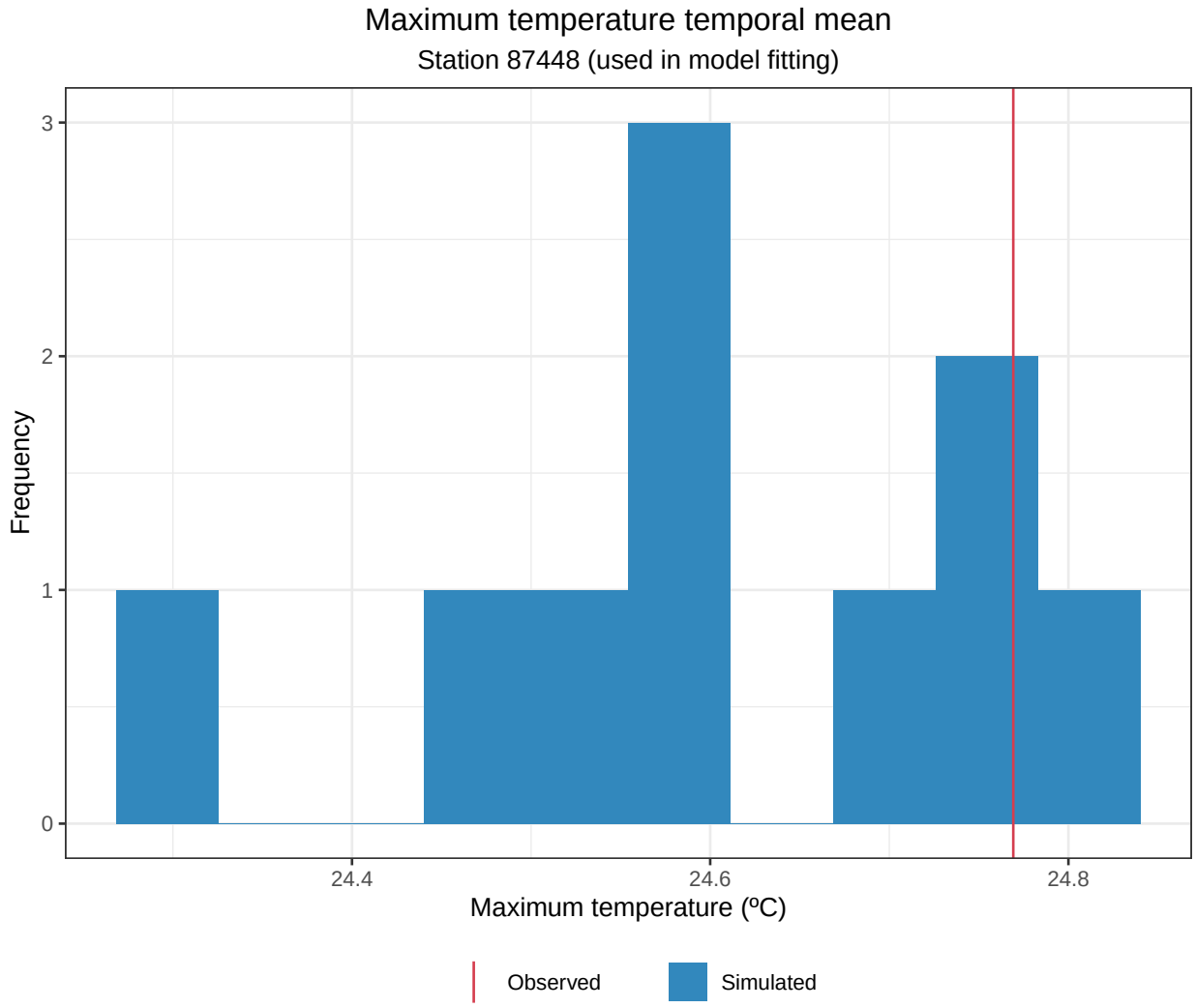
Esta figura muestra otra medida de dispersión, en este caso, el desvío absoluto mediano (DMA o MAD, por sus siglas en inglés). Al igual que en el caso anterior, las cajas corresponden al MAD calculado a partir de los datos observados y las cajas a las distintas realizaciones. El buen acuerdo entre las cajas y los puntos confirma que el generador captura la variabilidad en los totales mensuales de lluvia.

6.2 Diagnósticos de temperatura máxima

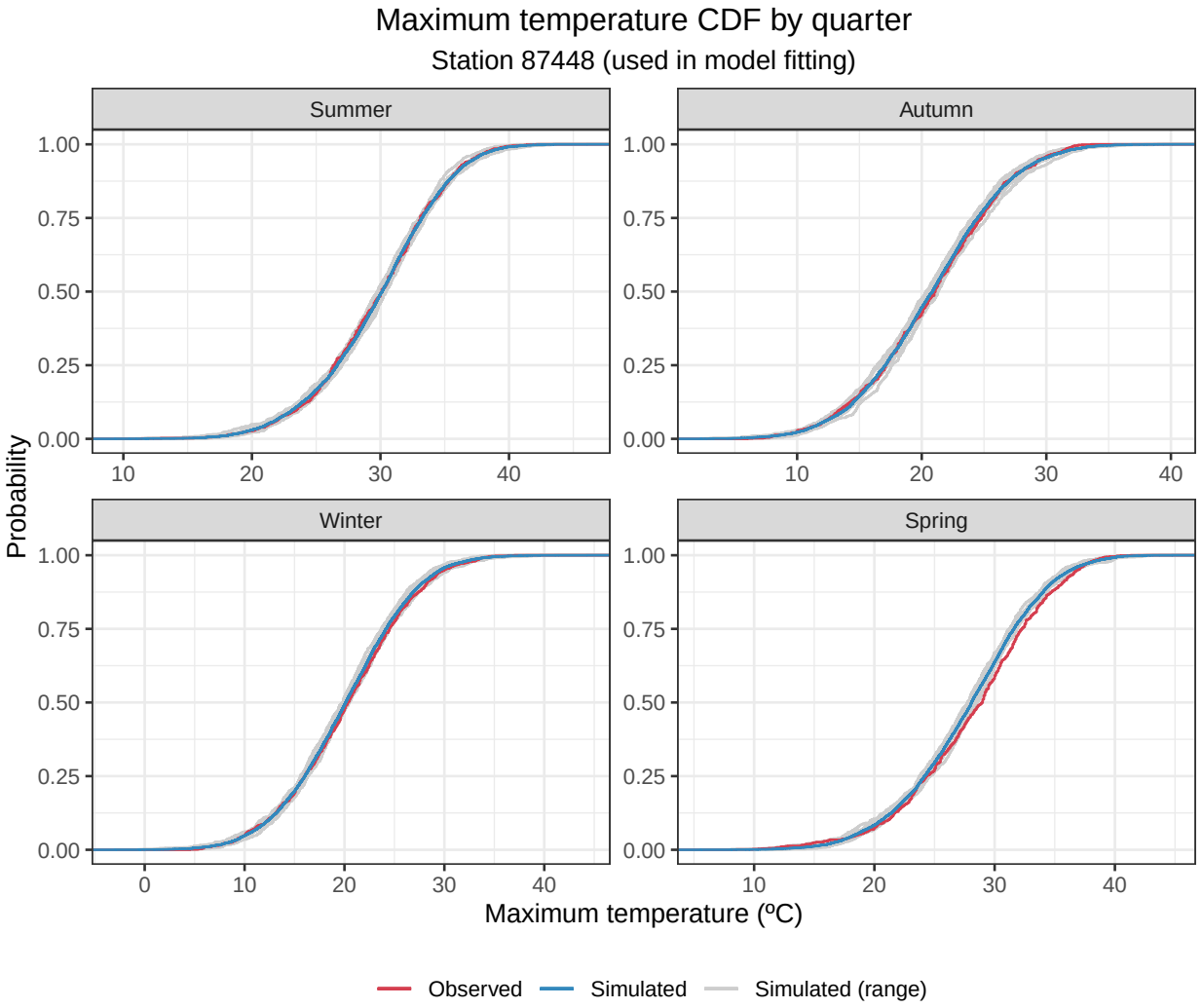
A continuación se mostrarán los distintos diagnósticos desarrollados para validar las series diarias de temperatura máxima.



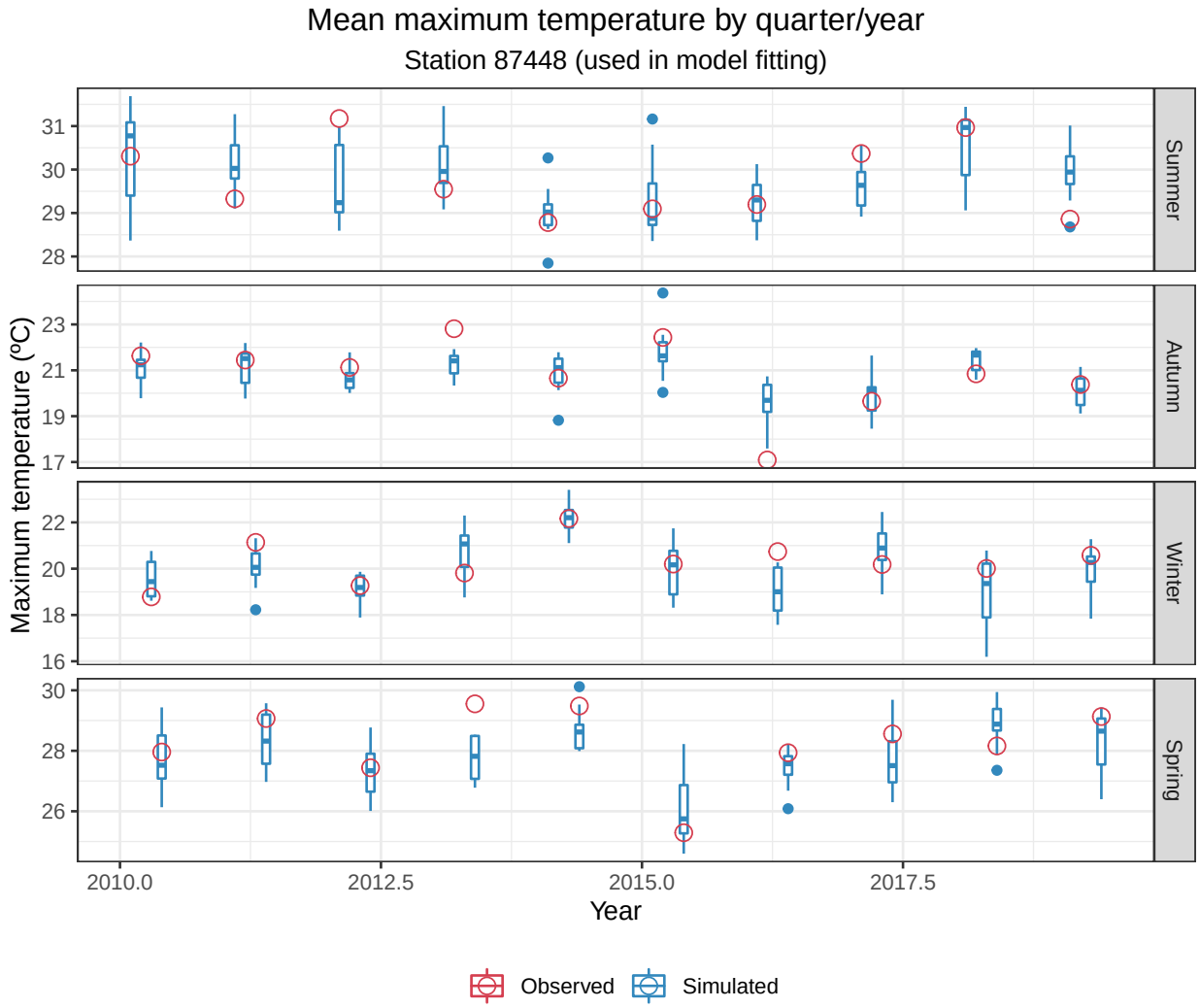
En esta figura se muestra una comparación de los cuantiles observados y generados de temperatura máxima diaria por trimestre. En el eje x se encuentran los cuantiles observados mientras que en el y, los generados. Los puntos azules corresponden a los puntos cuyo valor se encuentra entre el percentil 1 y 99 y los rojos a los menores y mayores a dicho percentiles, respectivamente. El objetivo de esta prueba es verificar que todos los puntos se encuentren sobre la recta 1:1. Se observa una muy buena concordancia para los cuatros trimestres incluso en los extremos de la distribución.



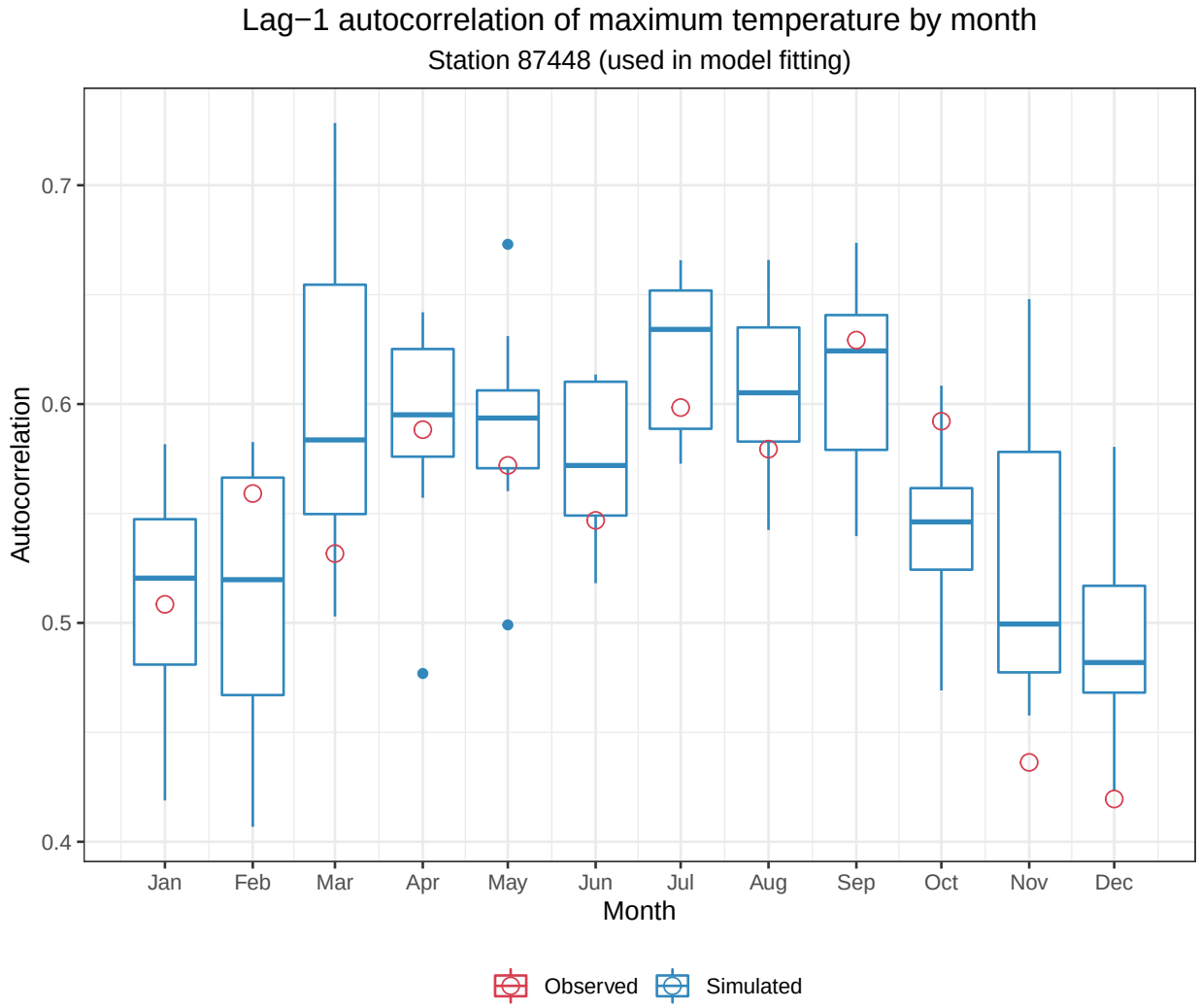
Esta figura compara la temperatura máxima media de la serie observada con la de las distintas realizaciones. La línea vertical roja corresponde al valor observado mientras que las barras a las distintas realizaciones. Se observa que la medias máximas de todos las realizaciones se encuentran en un rango muy estrecho del orden de décimas de grado centígrado y que la media histórica está muy próxima a la media de las distintas realizaciones.



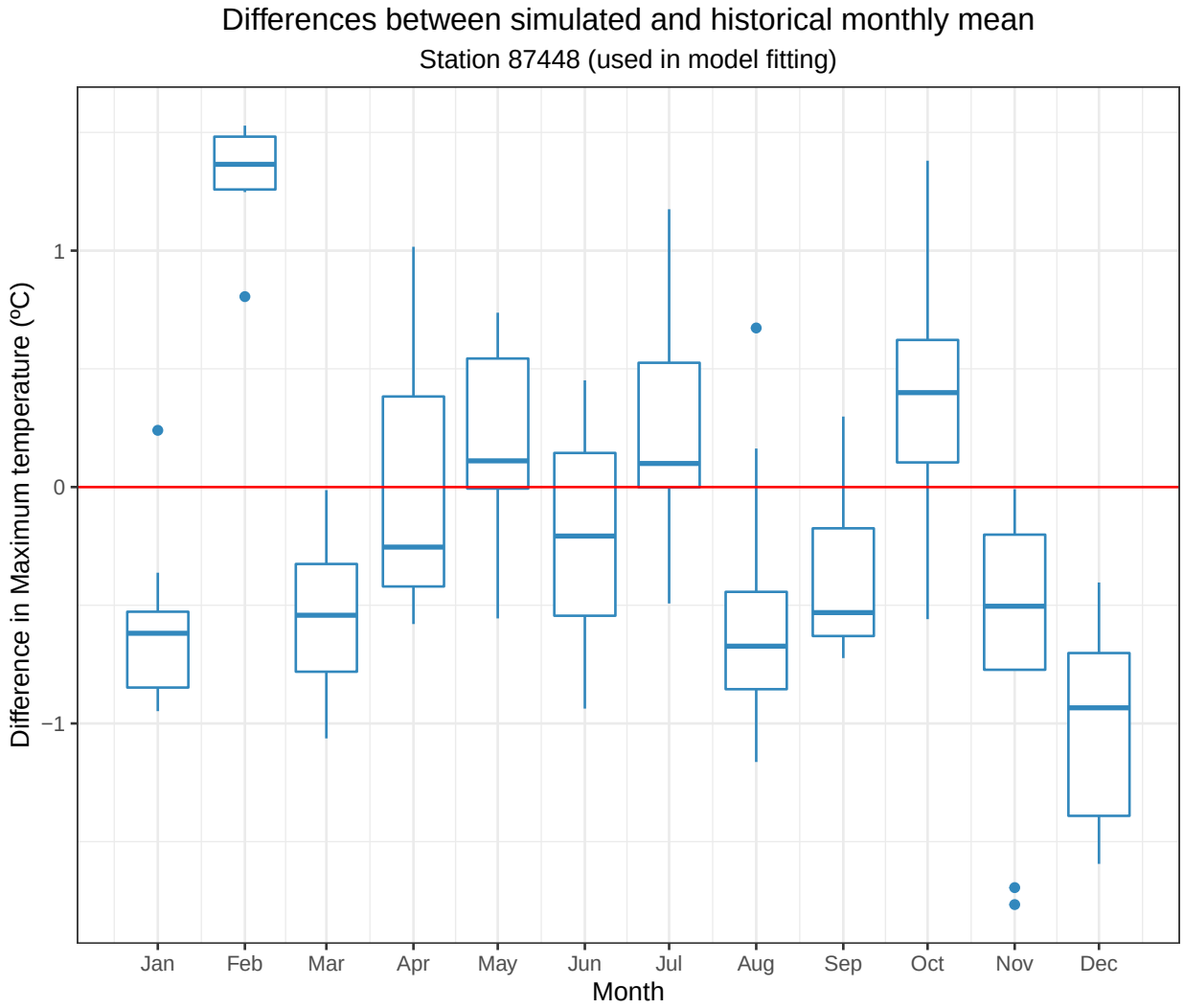
Esta figura muestra la probabilidad acumulada de la temperatura máxima por trimestre del año. La línea roja corresponde a la probabilidad observada mientras que las azules, a las distintas realizaciones. Se observa que las líneas azules envuelven perfectamente a la línea roja por lo que la distribución probabilística es casi idéntica.



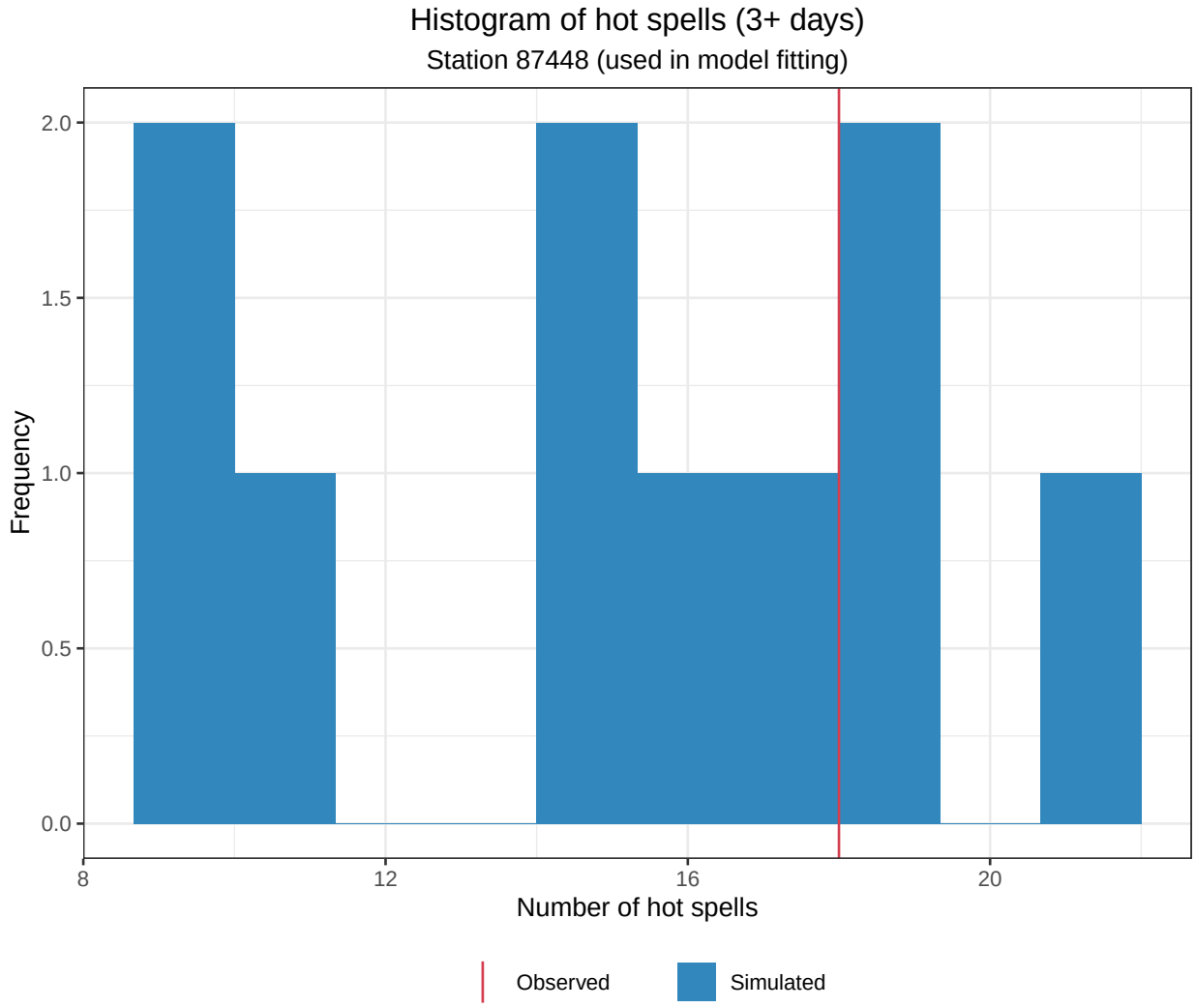
Esta figura compara la variabilidad de la temperatura máxima media por trimestre a lo largo del tiempo. Los puntos rojos corresponden a las medias observadas mientras que las cajas a las distintas realizaciones. Se observa que las cajas siguen el comportamiento de los datos observados aún en años extremos.



Esta figura compara la autocorrelación de los valores con el lag -1 de la variable. Es decir, compara la temperatura máxima diaria del día **t** con la temperatura máxima del día **t-1**. Los puntos rojos corresponden a la autocorrelación observada mientras que las cajas a las distintas realizaciones. Se observa un patrón anual muy marcado con autocorrelaciones más altas en invierno que en verano. Este diagnóstico es muy importante para verificar si las potenciales rachas cálidas serán bien reflejadas o no ya que si un día es muy cálido, el siguiente tenderá a serlo también.



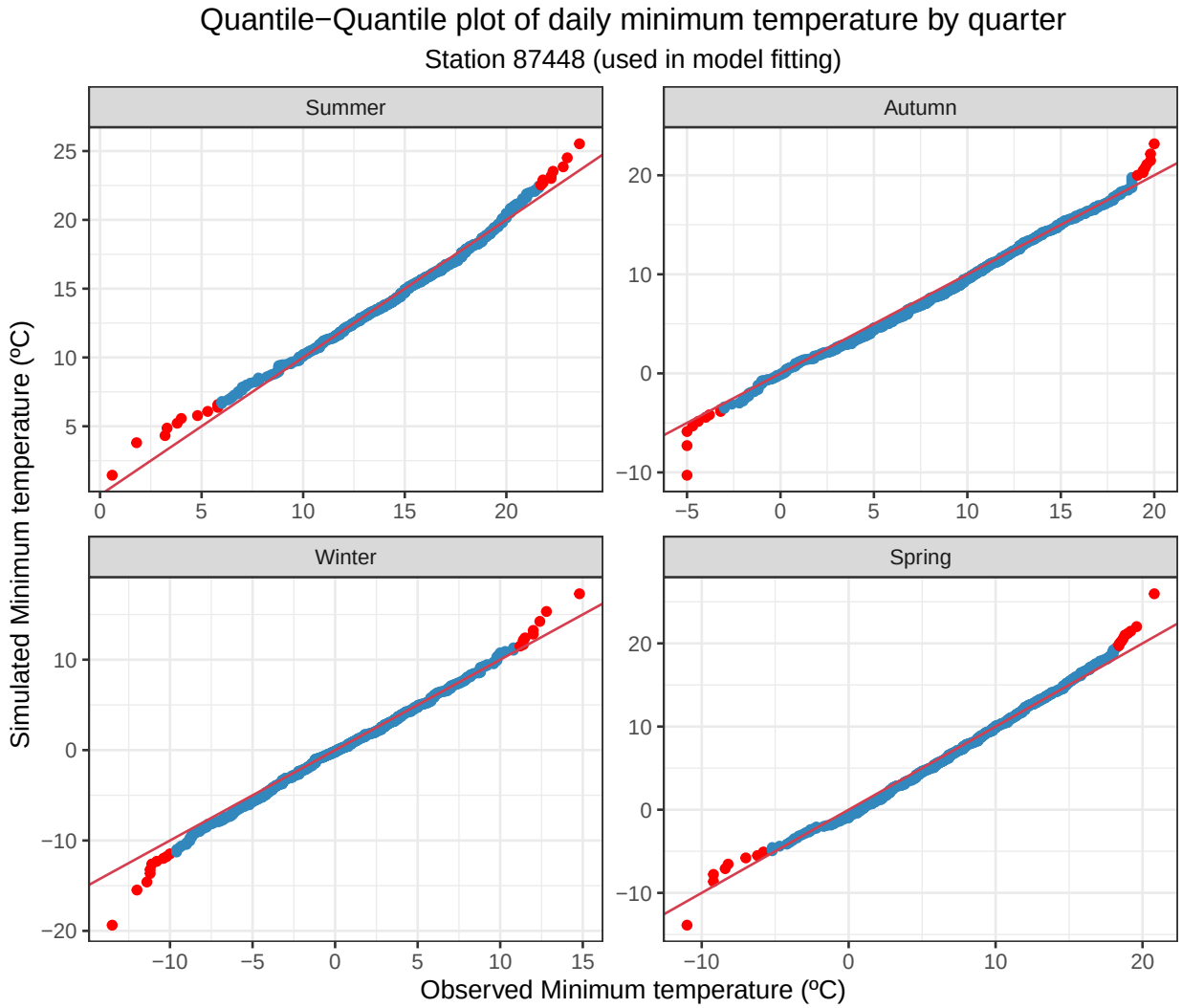
Esta figura compara la media mensual calculada a partir de la series históricas con las distintas realizaciones para cada mes del año. La línea roja corresponde a una diferencia de 0 lo que quiere decir que la media observada y generada son iguales. Se observa que para casi todos los meses las cajas se encuentran sobre la línea roja a excepción de los meses de junio y agosto donde se observa un leve sesgo cálido de menos de 0.25 °C.



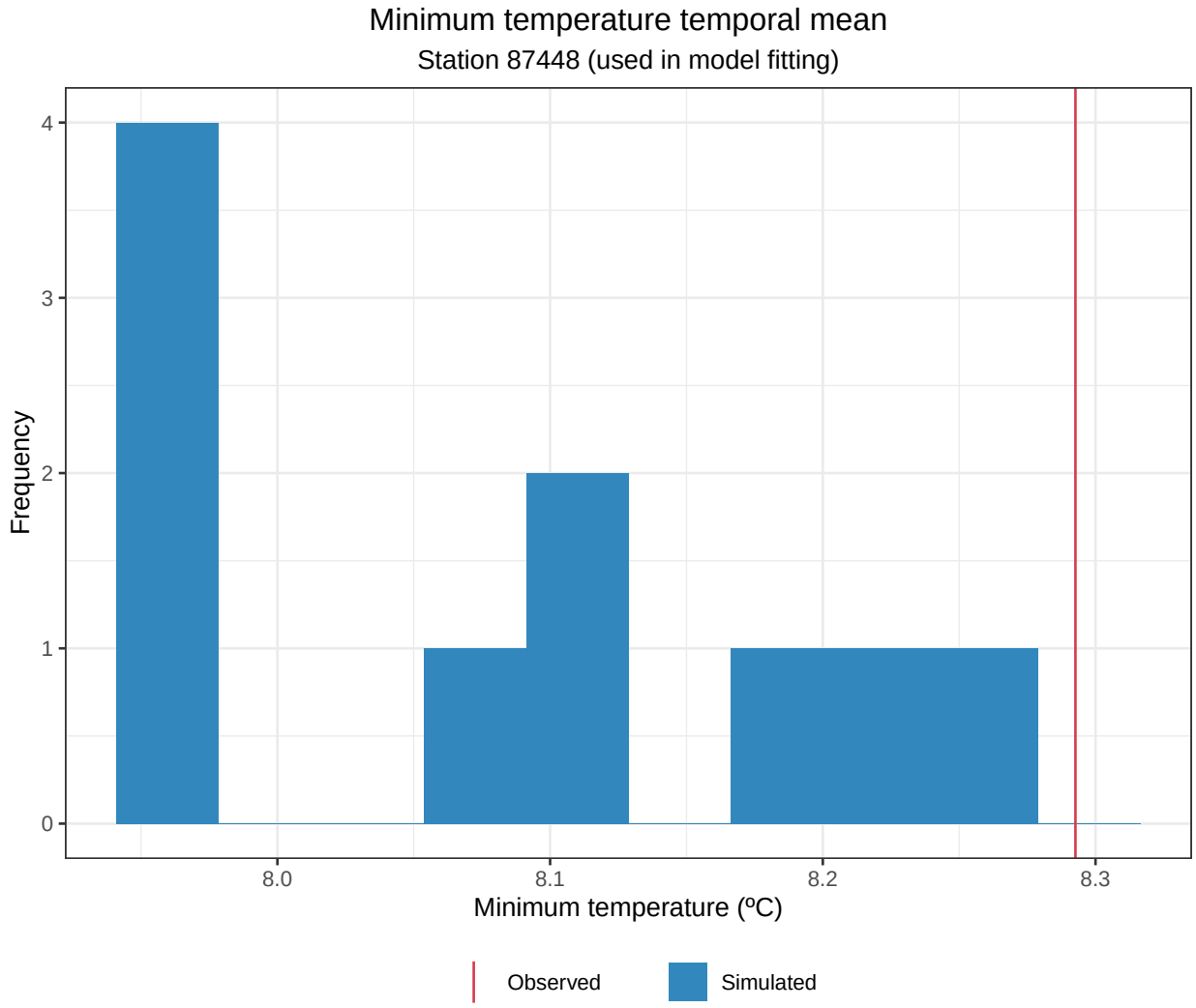
Esta figura compara la distribución de la cantidad de rachas cálidas generadas con las observadas en el registro histórico. La línea roja corresponde a la cantidad de períodos cálidos totales en el registro mientras que las barras a las distintas realizaciones. Se observa que hay una subestimación en la cantidad total de rachas cálidas para esta estación.

6.3 Diagnósticos de temperatura mínima

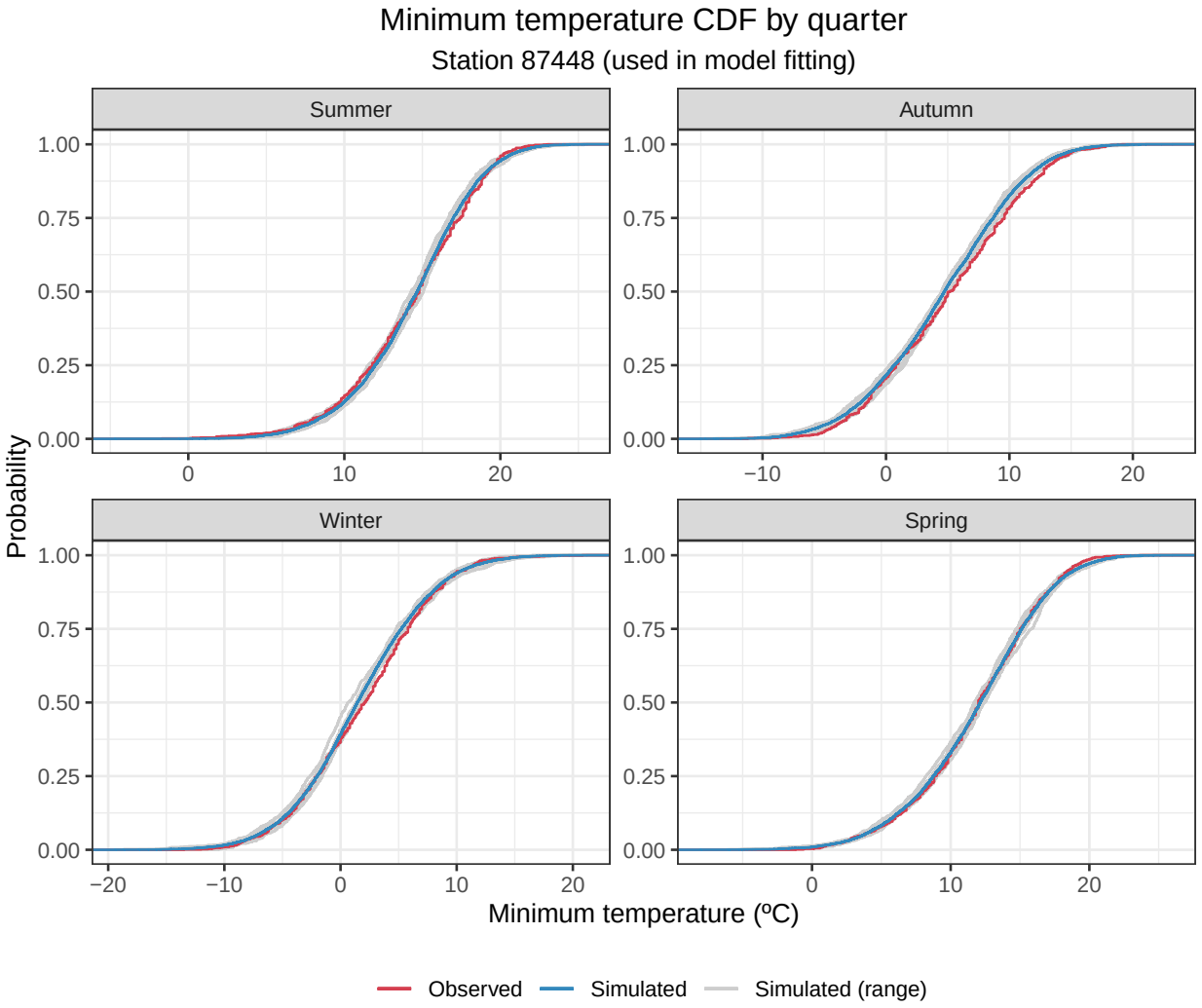
A continuación se mostrarán los distintos diagnósticos desarrollados para validar las series diarias de temperatura mínima



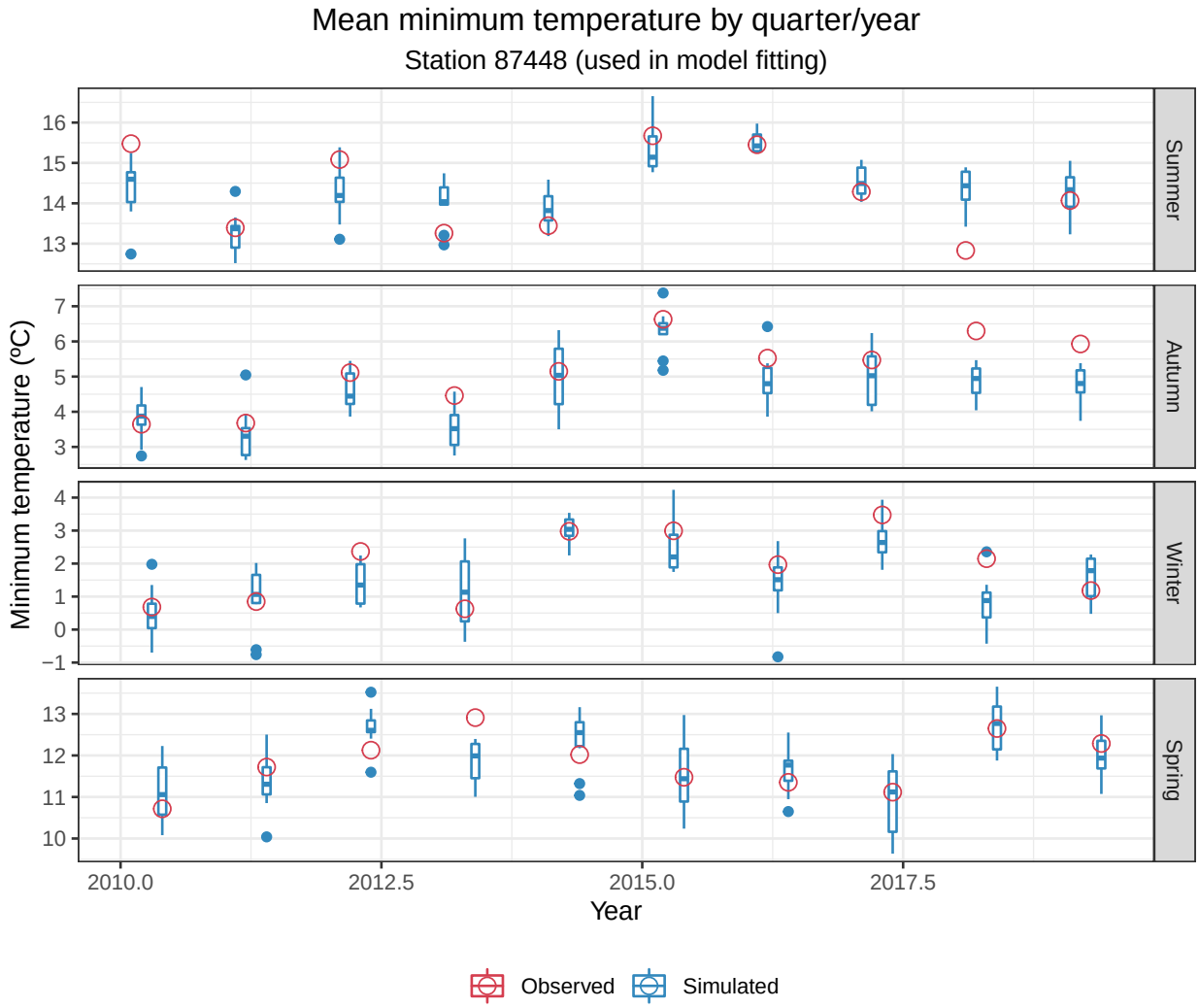
En esta figura se muestra una comparación de los cuantiles observados y generados de temperatura mínima diaria por trimestre. En el eje x se encuentran los cuantiles observados mientras que en el y, los generados. Los puntos azules corresponden a los puntos cuyo valor se encuentra entre el percentil 1 y 99 y los rojos a los menores y mayores a dichos percentiles, respectivamente. El objetivo de esta prueba es verificar que todos los puntos se encuentren sobre la recta 1:1. Se observa una muy buena concordancia para los cuatro trimestres incluso en los extremos de la distribución. Solo algunos puntos son subestimados en el extremo inferior de la distribución.



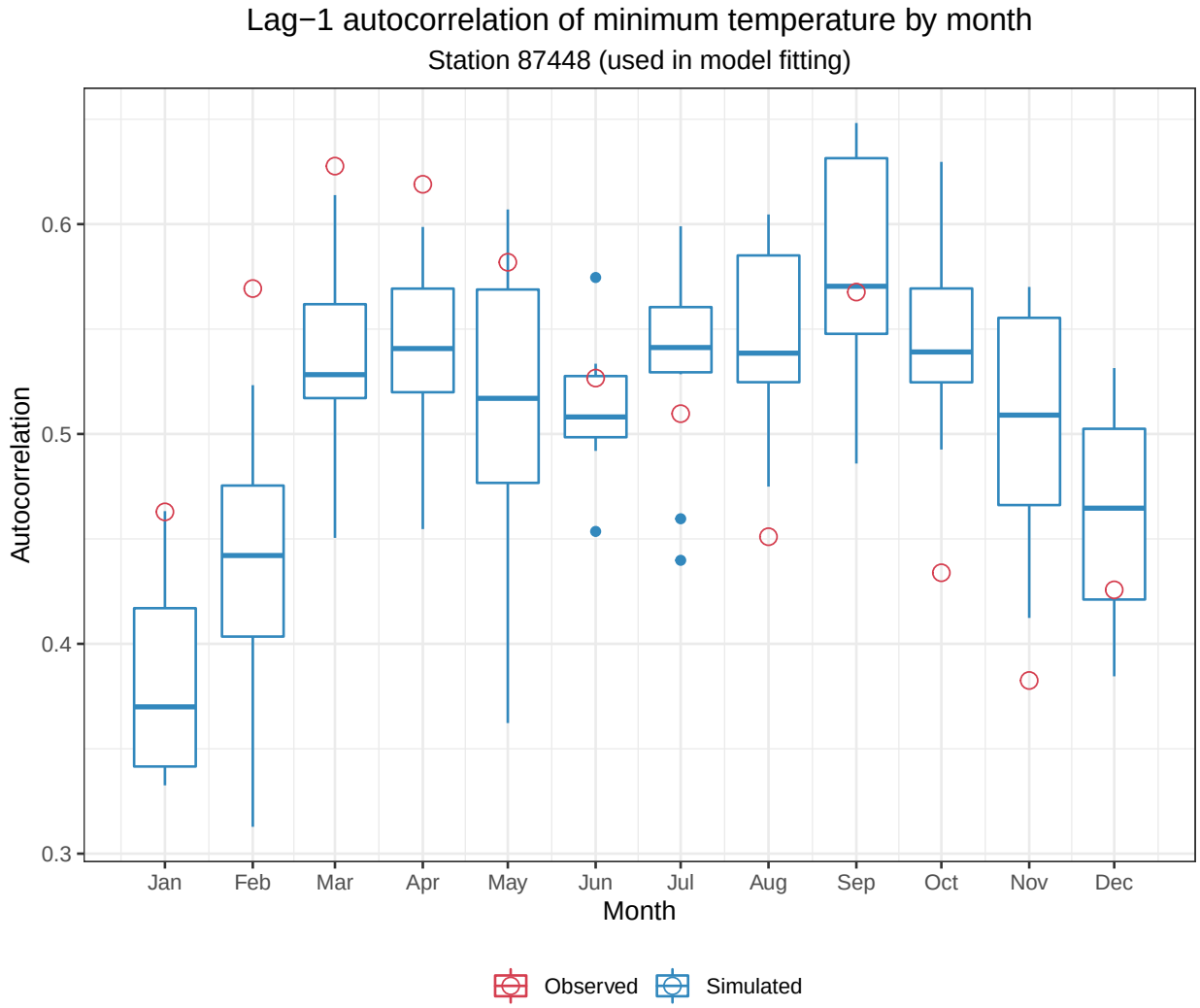
Esta figura compara la temperatura mínima media de la serie observada con la de las distintas realizaciones. La línea vertical roja corresponde al valor observado mientras que las barras a las distintas realizaciones. Se observa que la medias mínimas de todos las realizaciones se encuentran en un rango muy estrecho del orden de dos décimas de grado centígrado. La diferencia entre la media observada el centro de la distribución de las medias mínimas generadas es de menos de 0.2 °C.



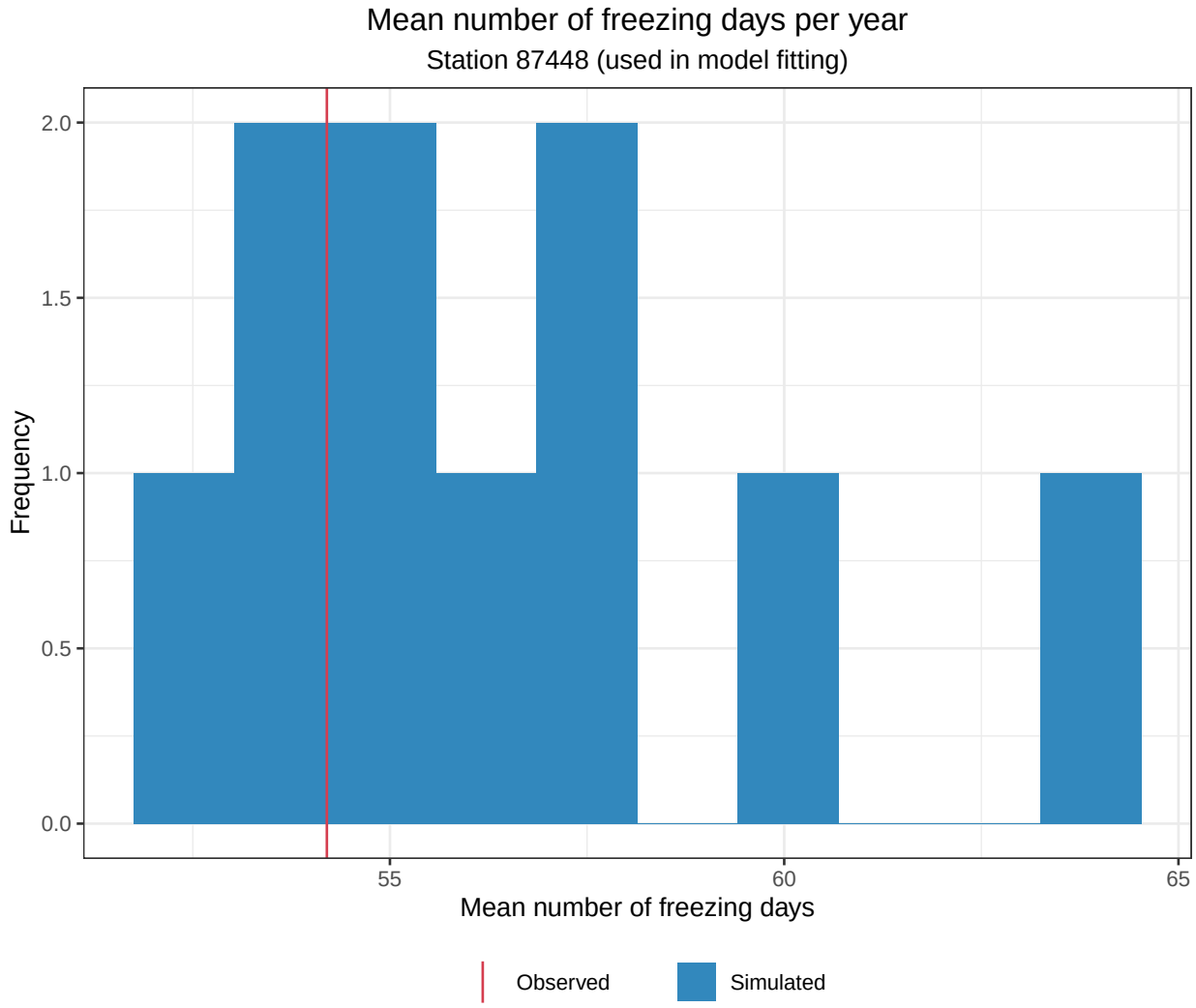
Esta figura muestra la probabilidad acumulada de la temperatura mínima por trimestre del año. La línea roja corresponde a la probabilidad observada mientras que las azules, a las distintas realizaciones. Se observa que las líneas azules envuelven perfectamente a la línea roja por lo que la distribución probabilística es casi idéntica.



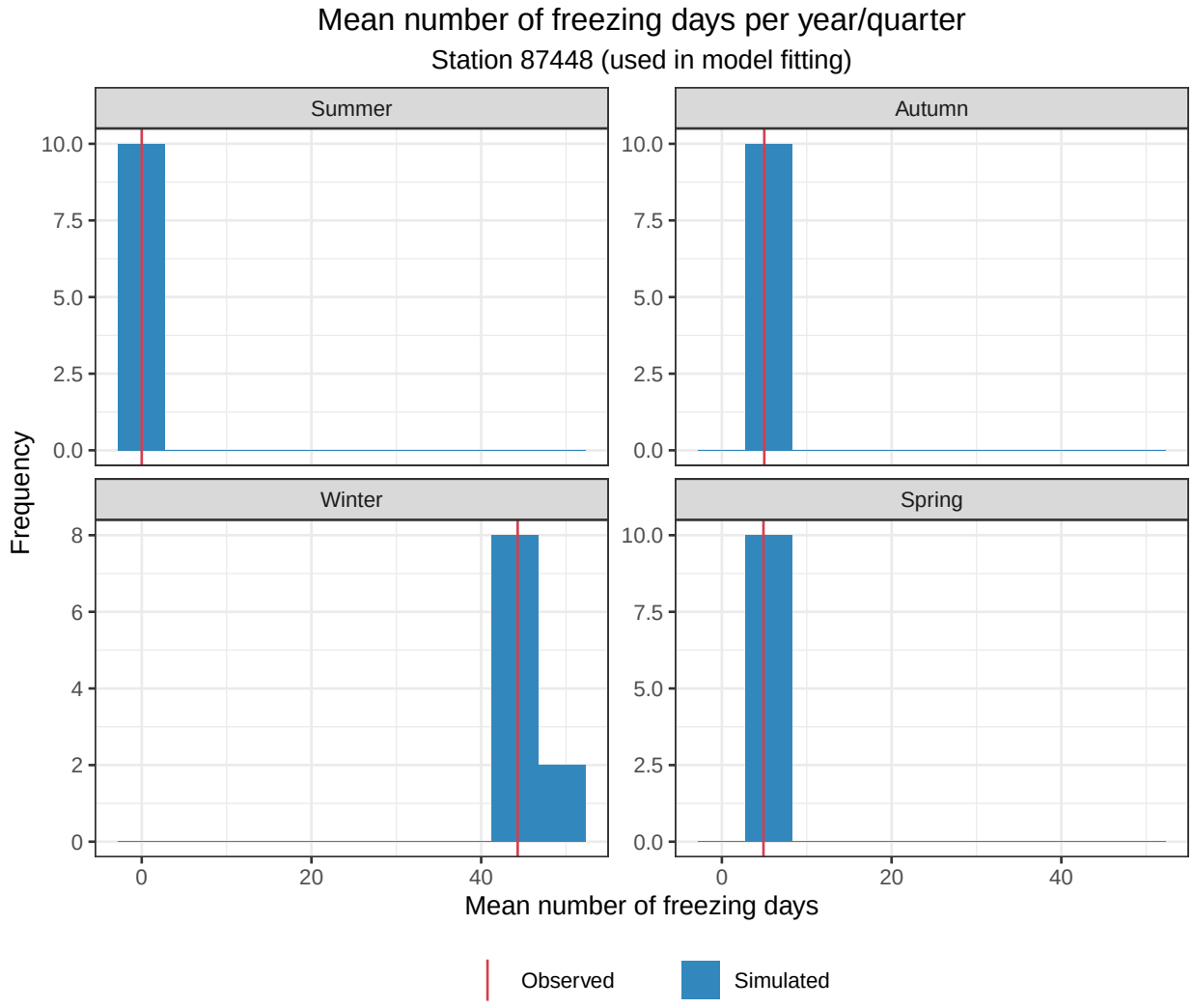
Esta figura compara la variabilidad de la temperatura mínima media por trimestre a lo largo del tiempo. Los puntos rojos corresponden a las medias observadas mientras que las cajas a las distintas realizaciones. Se observa que las cajas siguen el comportamiento de los datos observados incluso en años con grandes variaciones frente a la media.



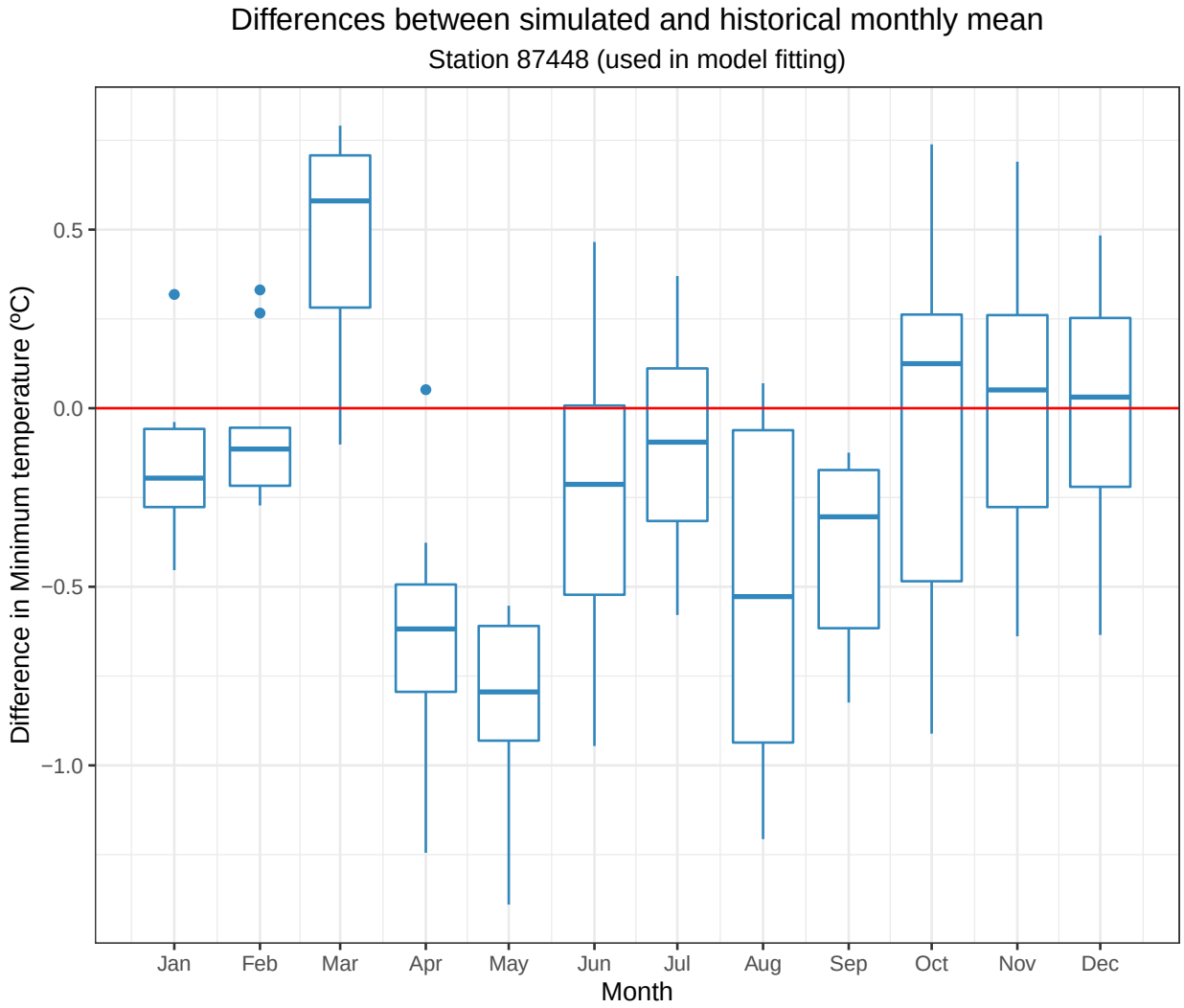
Esta figura compara la autocorrelación de los valores con el lag -1. Es decir, compara la temperatura mínima diaria del día **t** con la temperatura mínima del día **t-1**. Los puntos rojos corresponden a la autocorrelación observada mientras que las cajas a las distintas realizaciones. Se observa un patrón anual muy marcado con autocorrelaciones más altas en otoño e invierno que en verano.



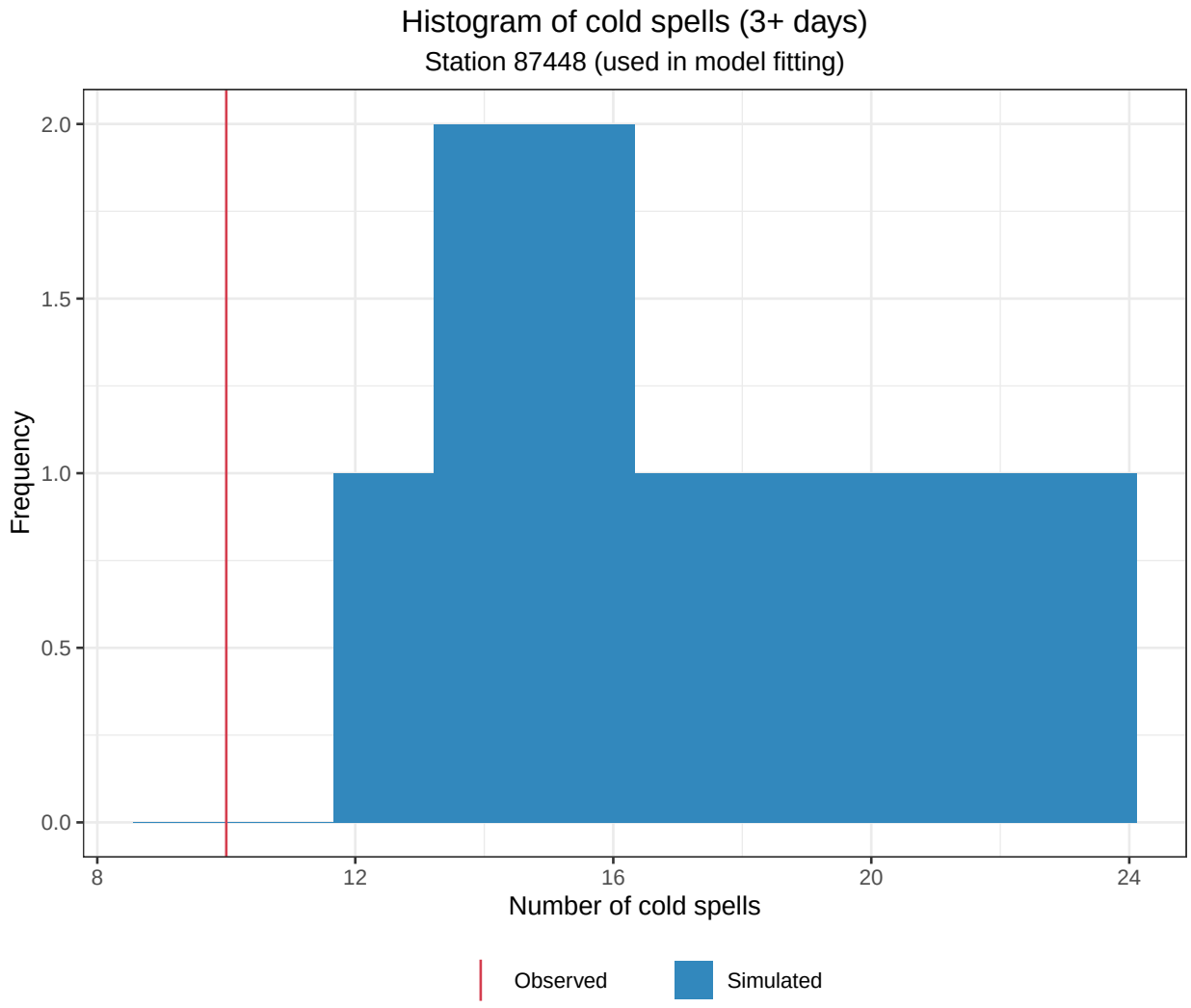
Esta figura compara la cantidad de días con heladas durante un año. La línea vertical roja corresponde a la cantidad media de heladas por año calculada a partir de los registros históricos mientras que las barras corresponden a las distintas realizaciones. El generador es capaz de simular muy bien los días con heladas ya que que la diferencia entre el valor observado y la media de la distribución de las series sintéticas es menor a un día.



Esta figura es similar a la anterior con la diferencia en que se divide el año en los distintos trimestres. Al igual que en el caso anterior, la cantidad de días con heladas generados es prácticamente idéntica a lo observado.



Esta figura compara la media mensual calculada a partir de la series histórica con las distintas realizaciones para cada mes del año. La línea roja corresponde a una diferencia de 0 lo que quiere decir que la media observada y generada son iguales. Se observa que durante los meses invernales hay un ligero sesgo cálido de 0.5 °C aproximadamente.

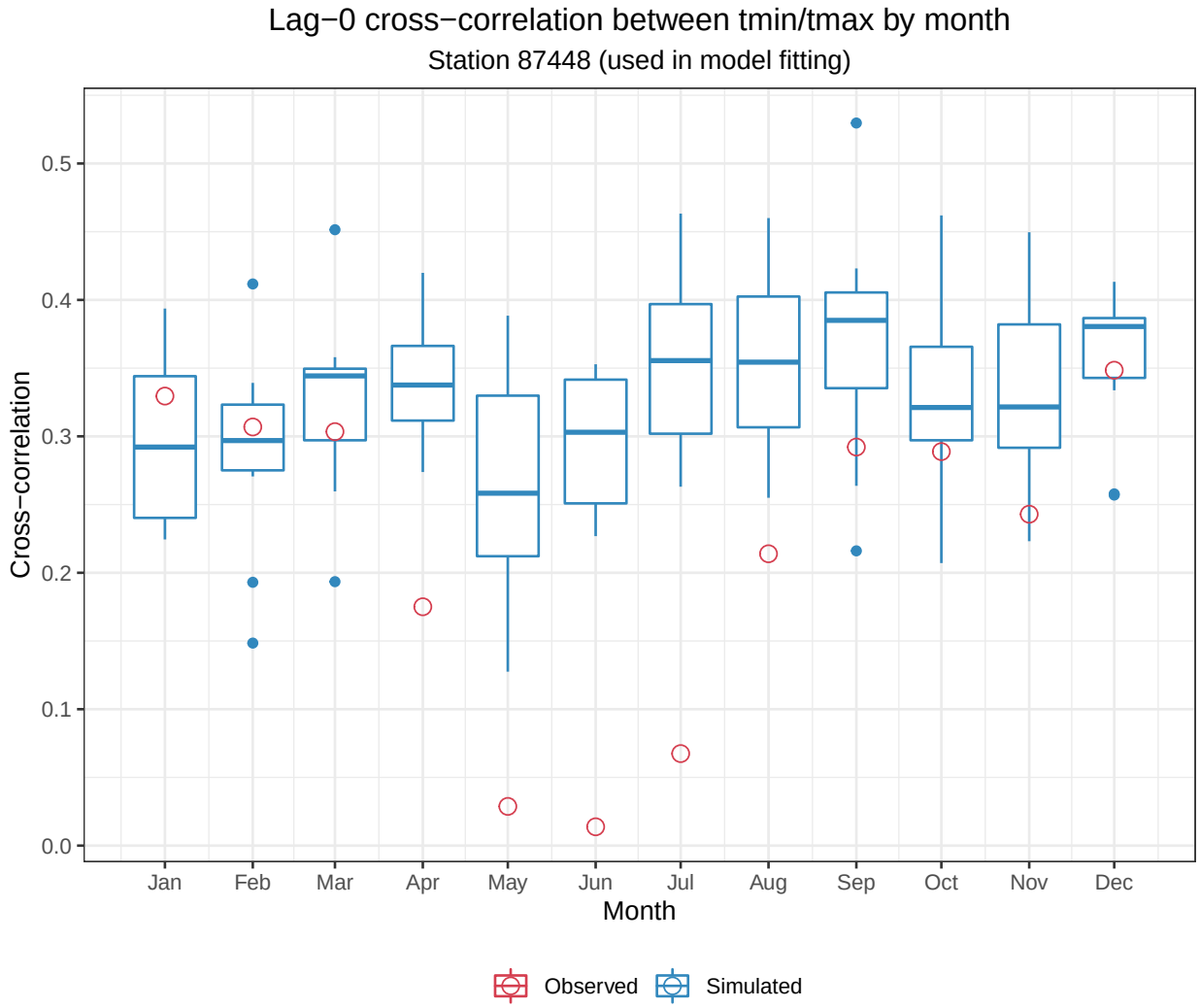


Esta figura compara la distribución de la cantidad de rachas frías generadas con las observadas en el registro histórico. La línea roja corresponde a la cantidad de períodos fríos totales en el registro mientras que las barras a las distintas realizaciones. Se observa que hay una subestimación en la cantidad total de rachas cálidas para esta estación.

6.4 Diagnósticos de auxiliares

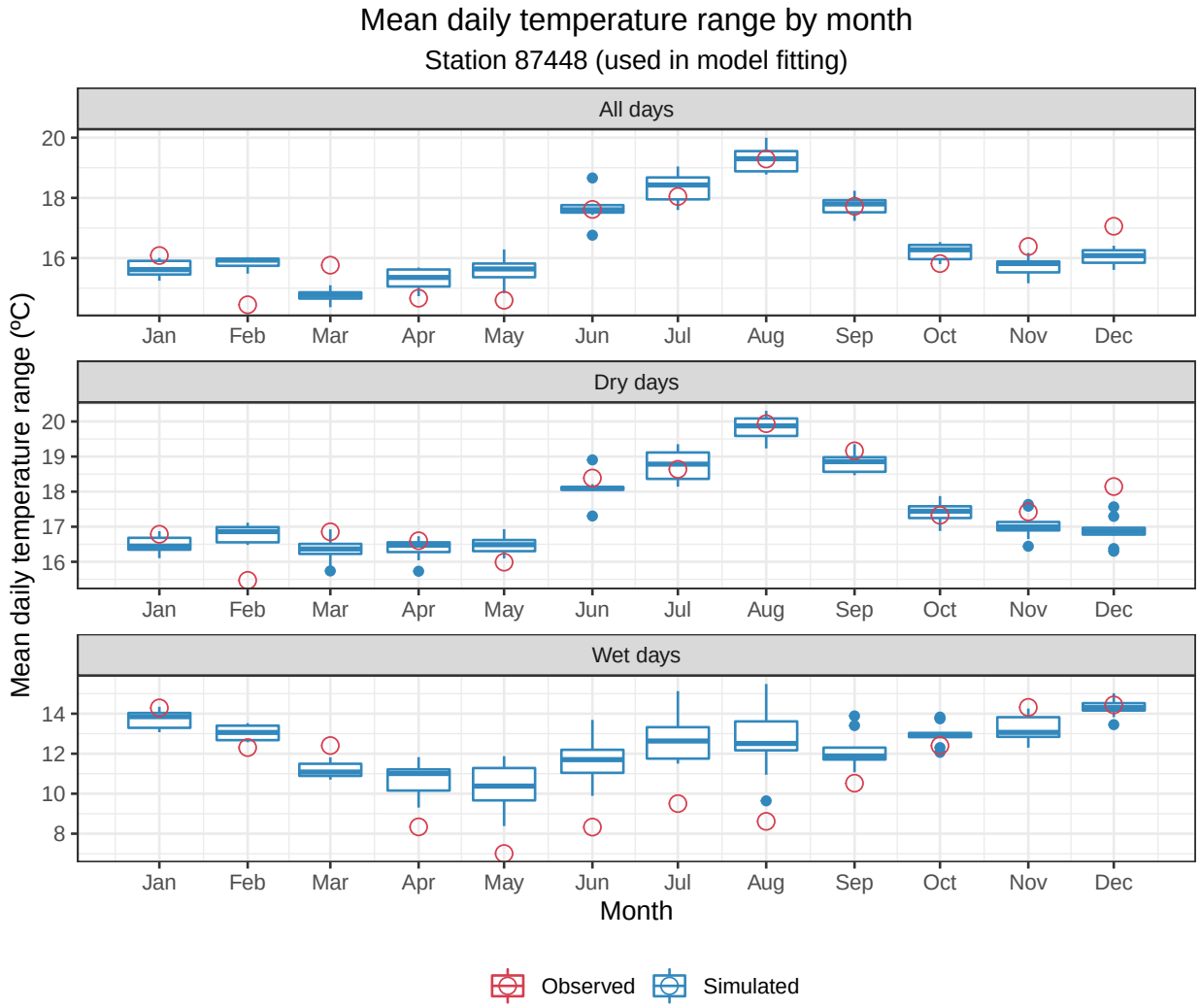
En esta sección se muestran de diagnósticos auxiliares centrados en la consistencia entre variables.

```
# Correlación cruzada entre las temperaturas máximas y mínimas por mes
plots[[station.id]][['other']][[1]]
```



Esta figura compara la variación mensual de la correlación cruzada entre las temperaturas máximas y mínimas. La autocorrelación cruzada evalúa si ambas variables varían al mismo tiempo y con que magnitud. Los puntos rojos corresponden a la correlación cruzada calculada a partir de los datos observados y las cajas a las distintas realizaciones. Se observa que se captura el patrón general aunque con una subestimación durante el invierno. Una baja autocorrelación implica grandes amplitudes térmicas ya que no varían conjuntamente.

```
# Temperatura media diaria mensual por tipo de día (seco o lluvioso)
plots[[station.id]][['other']][[2]]
```



Esta figura compara la temperatura media para los distintos meses del año considerando el tipo de día, es decir, si llovió o no. Los puntos rojos corresponden al valor observado mientras que las cajas a las distintas realizaciones. Se observa que cuando se consideran todos los días o sólo los días secos la concordancia es muy buena. Para los días lluviosos, en cambio, ésta sigue siendo buena para los meses estivales pero existe un marcado sesgo cálido en el invierno. Esto podría deberse a que en esta estación hay muy poco días lluviosos en invierno por lo que no son suficientes para que el modelo los capture.